Universidad de Cuenca

Facultad de Ingeniería

Escuela Informática

"PROCESO DE TRANSFORMACIÓN Y VISUALIZACIÓN DE DATOS SEMÁNTICOS A DATOS DIMENSIONALES."

Tesis de grado previa a la obtención
del título de Ingeniero de Sistemas

Autores:

Angélica Azucena Cárdenas Guzhñay.
CI: 0106527229
Viviana Lucía Proaño Barros.
CI: 1716341795

Director:

Ing. Víctor Hugo Saquicela Galarza, PhD
CI: 0103599577

Abril 2017


UNIVERSIDAD DE CUENCA

Resumen

La web semántica ha tenido gran acogida para ser una alternativa en el almacenamiento de datos, ya que a diferencia de las bases relacionales, la web semántica permite que los datos tengan significado tanto para las personas como para los ordenadores. El proyecto REDI, posteriormente explicado, utiliza esta forma de almacenamiento, los datos son anotados semánticamente sobre una ontología que sirve para almacenar datos de publicaciones y autores que son el objetivo de dicho proyecto, estos datos están sobre una estructura semántica en formato RDF. Uno de los problemas que se ha identificado al tener los datos sobre la estructura en RDF es el tiempo de respuesta al ejecutar consultas sobre el grafo central que almacena la información, ya que estas son estáticas, el lenguaje utilizado para acceder a los datos semánticos es el lenguaje SPARQL. En este trabajo de titulación se pretende utilizar una segunda alternativa para almacenamiento de datos semánticos, esta alternativa pretende dar solución al problema de los tiempos de respuesta utilizando estructuras multidimensionales, para llevar a cabo esta transformación de estructuras de almacenamiento la W3C propone el vocabulario de Cubo de Datos que permite que los datos sean almacenados bajo una estructura semántica multidimensional, permitiendo de esta manera que las consultas sean más dinámicas y reduciendo así los tiempos de respuesta cuando se accede a los datos. Teniendo los datos almacenados en el Cubo de Datos nace la necesidad de visualizarlos, los Cubos de Datos tienen la ventaja de que la estructura permite que estos sean visualizados estadísticamente, por lo tanto para la visualización se utiliza, en este caso, herramientas que son específicas para este propósito, por ejemplo OpenCube Toolkit.

Palabras Clave: Cubo de Datos, REDI, Visualización de Datos, Medios de almacenamiento de Datos, OpenCube Toolkit.

Abstract

The semantic web has been well received to be an alternative in the storage of data, unlike the relational databases, the semantic web allows the data to have meaning for both people and computers. The REDI project, later explained, uses this form of storage, the data are annotated semantically on an ontology that serves to store data of publications and authors that are the objective of this project, these data are on a semantic structure in RDF format. One of the problems that has been identified when having the data about the structure in RDF is the response time when executing queries on the central graph that stores the information, since these are static, the language used to access the semantic data is The SPARQL language. This thesis intended to use a second alternative for storing semantic data, also this alternative aims to solve the problem of response times using multidimensional structures to carry out this transformation of storage structures the W3C proposes the vocabulary of Cube Of Data that allows the data to be stored under a multidimensional semantic structure, allowing in this way that the queries are more dynamic and thus reducing the response times when accessing the data. Taking the data stored in the Data Cube is born the need to visualize them, Data Cubes have the advantage that the structure allows them to be displayed statistically, therefore for the visualization is used, in this case, tools that are specific for this purpose, for example OpenCube Toolkit.

Keywords: Data Cube, REDI, Data Visualization, Data Storage Media, OpenCube Toolkit.

Índice general

Índice de figuras

Índice de cuadros

Universidad de Cuenca
Clausula de derechos de autor

Angélica Azucena Cárdenas Guzhñay, autora de la tesis "Proceso de Transformación y Visualización de Datos Semánticos a Datos Dimensionales", reconozco y acepto el derecho de la Universidad de Cuenca, en base al Art. 5 literal c) de su Reglamento de Propiedad Intelectual, de publicar este trabajo por cualquier medio conocido o por conocer, al ser este requisito para la obtención de mi título de Ingeniera de Sistemas. El uso que la Universidad de Cuenca hiciere de este trabajo, no implicará afección alguna de mis derechos morales o patrimoniales como autora.

Cuenca, 10 de abril del 2017.

Angélica Azucena Cárdenas Guzhñay

C.I: 0106527229

Universidad de Cuenca
Clausula de propiedad intelectual

Angélica Azucena Cárdenas Guzhñay, autora de la tesis "Proceso de Transformación y Visualización de Datos Semánticos a Datos Dimensionales", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autora.

Cuenca, 10 de abril del 2017.

Angélica Azucena Cárdenas Guzhñay

C.I: 0106527229

Universidad de Cuenca
Clausula de propiedad intelectual

Viviana Lucía Proaño Barros, autora de la tesis "Proceso de Transformación y Visualización de Datos Semánticos a Datos Dimensionales", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autora.

Cuenca, 10 de abril del 2017

Viviana Lucía Proaño Barros

C.I: 1716341795

Universidad de Cuenca
Clausula de derechos de autor

Viviana Lucía Proaño Barros, autora de la tesis "Proceso de Transformación y Visualización de Datos Semánticos a Datos Dimensionales", reconozco y acepto el derecho de la Universidad de Cuenca, en base al Art. 5 literal c) de su Reglamento de Propiedad Intelectual, de publicar este trabajo por cualquier medio conocido o por conocer, al ser este requisito para la obtención de mi título de Ingeniera de Sistemas. El uso que la Universidad de Cuenca hiciere de este trabajo, no implicará afección alguna de mis derechos morales o patrimoniales como autora

Cuenca, 10 de abril del 2017

Viviana Lucía Proaño Barros

C.I: 1716341795

Agradecimientos

El presente trabajo de titulación es un logro obtenido gracias al apoyo de familiares, amigos y compañeros, quienes con su amistad y apoyo ayudaron a que este trabajo de titulación culmine con éxito, además queremos agradecer a nuestros docentes quienes han compartido con nosotras sus conocimientos para poder seguir en nuestra carrera profesional.

Queremos agradecer de manera muy especial a nuestro asesor de tesis, Ingeniero Víctor Saquicela, por su esfuerzo, dedicación y por aportar con sus conocimientos para llevar a cabo este trabajo.

Agradecemos también al equipo del proyecto OpenCube Toolkit, con especial mención Evangelos Kalampokis, quien nos ha brindado su apoyo durante la fase de visualización del trabajo de titulación, gracias por la confianza y el apoyo, el cual nos permitió culminar con el proyecto.

Dedicatoria

Dedico este trabajo de titulación a mis padres, Angel y Laura, que con su amor, apoyo incondicional y ejemplo han sabido guiarme en todos los aspectos de mi vida.

A mi hermana Nancy, que siempre ha sido mi apoyo para seguir adelante, por ayudarme con sus conocimientos y por el buen ejemplo que siempre ha sido para mí.

A mi familia, que de una u otra forma me han apoyado para perseverar y poder cumplir mis metas.

Angélica

Dedicatoria

Primeramente, dedico este trabajo de titulación a Dios, quien se ha manifestado de varias maneras en mi vida y ha permitido este nuevo logro y a mi madre Santisima, la Virgen María, quien me ha acompañado en este caminar.

A mis queridos abuelitos, Luis y Olga quienes con mucho amor, ejemplo y firmeza, me han sabido inculcar su fe y valores que han sido los pilares fundamentales en mi vida.

A mi querida madre, Lucía, por su amor y ejemplo de lucha, perseverancia y trabajo constante que me han dado la fortaleza para jamás rendirme y lograr lo que me he propuesto.

A mi querido padre, Marco, por su amor, preocupación, confianza y apoyo que me ha brindado.

A mi querida hermana, Tatiana, por ser mi apoyo y ejemplo de superación de adversidades con alegría y positivismo.

También les dedico a todos mis compañeros que durante todos estos años han venido compartiendo momentos y anécdotas que perdurarán durante toda la vida.

Viviana

Capítulo 1

Introducción

En este capítulo se describe la problemática que ha dado lugar al desarrollo del presente trabajo de titulación, además se incluye el alcance y los objetivos del mismo.

1.1. Panorama General

La presente tesis está relacionada con el incremento masivo de datos en la web, los datos en formato RDF han sido por mucho tiempo una alternativa de consulta rápida hacia estos datos. CEDIA se encuentra actualmente trabajando en el desarrollo de un proyecto que servirá para consultar publicaciones de los docentes de las Universidades que pertenecen a la Red CEDIA. El problema identificado se refiere a que los datos cada vez van en aumento, haciendo que los tiempos de consultas sean cada vez mayores, por lo tanto, se ha optado cambiar la estructura de datos actual (RDF) a una estructura de datos multidimensional (Cubo de Datos) haciendo que los datos estén organizados por dimensiones, atributos y medidas, esto hace que el acceso a los datos sea más rápido debido a que se ignoran de cierta manera las dimensiones de datos no consultados. Se desarrollará la estructura de los datos y la visualización estadística de los mismos, haciendo uso del sistema OpenCube Toolkit, el cual será modificado para cumplir con el objetivo planteado.

1.2. Identificación del problema

El manejo de datos e información es un tema que en los últimos años ha incrementado tanto en investigación como en la aplicación para estrategias de negocios. Las bodegas de datos integran y consolidan grandes cantidades de información provenientes de diferentes fuentes de una organización; la organización de estos datos permite el análisis de: patrones en la información, comportamientos a lo largo del tiempo, integración de información, planteamiento de consultas, detectar casos anómalos y análisis de grandes volúmenes de información de manera rápida y confiable.

En la actualidad, el manejo de grandes volúmenes de datos hace que su manipulación sea cada vez más compleja, es decir, el problema se identifica en el incremento de las fuentes de datos, por lo tanto surge la necesidad de un mayor procesamiento para la obtención de información. Los tiempos de acceso y extracción de datos incrementan considerablemente debido a que se recorre todas las tripletas del RDF almacenadas en los repositorios web. A pesar que estos datos estén organizados de acuerdo a una ontología, aún se requiere un nivel más alto de organización. La extracción se realiza mediante SPARQL, siendo esta la única alternativa de consulta sobre el modelo.

Con el propósito de reducir los tiempos de acceso a los datos, varias organizaciones optan por implementar Cubos de Datos, debido a que el diseño, y la construcción permite escalar progresivamente hacia una arquitectura de almacenamiento de tipo Data Warehouse. Esta arquitectura mejora el tiempo de acceso y extracción de información por lo que permite mejorar las estrategias en el mercado.

1.3. Justificación del Problema

El incremento de datos en la web es cada vez mayor, por lo que se deben buscar alternativas más livianas para acceder a los datos, la web semántica es una alternativa, pero para consultar

fuentes de datos que están en RDF el tiempo de acceso está relacionado con la cantidad de datos que se encuentran en el repositorio web.

Las organizaciones necesitan tener rápido acceso a los datos y que estos sean confiables y oportunos, para saber en el momento exacto como está el funcionamiento de su organización, otra alternativa son los modelos multidimensionales, los cuales ayudan a reducir tiempos de acceso a las fuentes de datos debido a su organización por valores, metadatos y enlaces; los cuales facilitan la exploración en el modelo. Para aprovechar las ventajas de los modelos multidimensionales, los datos que se encuentran en RDF deben pasar por un proceso de transformación hacia un modelo de Cubo de Datos, con el propósito de optimizar el tiempo de acceso y extracción de datos. Principalmente los datos en RDF deben ser analizados para definir la estructura de las dimensiones y las medidas correspondientes al Cubo de Datos; el mapeo es la etapa principal de esta transformación debido a que se organiza de acuerdo al modelo del Cubo de Datos que se ha definido, posteriormente estos datos organizados multidimensionalmente deben ser presentados de forma estadística al usuario para lo cual existen ciertas herramientas que ayudan a los desarrolladores a presentar la información de un Cubo de Datos. Existen herramientas de visualización tales como CubeViz, Payola, OpenCube Toolkit, entre otros.

La visualización de la información en formato RDF, se ha convertido en un punto de interés para el usuario, sin embargo, actualmente resulta algo complejo tanto para el usuario como para el desarrollador. Para el usuario se vuelve tedioso tener que interpretar la información presentada, esto requiere mayor tiempo por parte del usuario, por otro lado, para el desarrollador es complejo presentar de forma rápida, organizada y concreta las especificaciones del usuario.

1.4. Alcance

Con la presente tesis se trabajará con información de artículos publicados por docentes de las Universidades que pertenecen a la red CEDIA. Se pretende mejorar la organización de los datos que son extraídos de varias fuentes como Scopus, Dblp, Microsoft Academics y Google

Scholar, los cuales son previamente tratados y almacenados en repositorios web en formato RDF.

Se realizará el análisis e implementación de un sistema que ayude en el proceso de transformación de datos en RDF hacía una estructura multidimensional. Primeramente se identifican los atributos, dimensiones y medidas que servirán para montar la estructura del Cubo de Datos, por lo tanto se realiza un mapeo sobre el RDF para identificar las entidades y asociarlas con el vocabulario de Cubo de Datos propuesto por la W3C, la estructura del Cubo de Datos se realiza en la herramienta Protege, posteriormente para instanciar el Cubo de Datos el sistema accede a Marmotta de donde serán extraídos los datos mediante una consulta SPARQL, estos datos obtenidos se almacenan dentro del Cubo de Datos previamente definido. El sistema dará como resultado un archivo en formato TURTLE el cual podrá ser visualizado estadísticamente utilizando una herramienta Open Source *OpenCube Toolkit*, la misma que será adecuada de acuerdo a los objetivos propuestos en este proyecto de titulación.

1.5. Objetivo General

Desarrollar un sistema que permita la transformación de datos semánticos a datos multidimensionales, los cuales serán visualizados de manera estadística.

1.6. Objetivos Específicos

Los objetivos específicos de esta tesis son:

1. Analizar modelos alternativos (multidimensionales).

2. Transformar datos RDF a RDF Cube.

3. Visualización de un RDF Cube.

1.7. Metodología aplicada

La metodología que se plantea para el presente proyecto de tesis se basa en las siguientes fases:

- Búsqueda bibliográfica: Esta fase básicamente consiste en la búsqueda de artículos científicos, revistas, libros, es decir, toda aquella fuente bibliográfica que pueda proveer información seria referente al tema de investigación.
- Clasificación de bibliografía: Esta segunda fase consiste en clasificar y filtrar la información recolectada del paso anterior, esto con el fin de desechar todo aquel material que no brinde información relacionada con los temas de investigación.
- Clasificación de bibliografía: Esta segunda fase consiste en clasificar y filtrar la información recolectada del paso anterior, esto con el fin de desechar todo aquel material que no brinde información relacionada con los temas de investigación.
- Investigación bibliográfica: Con la bibliografía ya clasificada se procede a obtener información y a relacionarla con los distintos puntos que se encuentran en el esquema de investigación.
    - En esta fase también se realizará procesos que permitan entender el mapeo de los metadatos con las ontologías (clases, propiedades, etc.), esto de forma manual.
- Redacción: Con la información obtenida en el paso anterior se procederá a redactar lo que sería el marco teórico de la investigación.
- Elección de herramientas: Con los conocimientos adquiridos se realiza una clasificación y comparación de las distintas herramientas existentes para el desarrollo de la aplicación de la presente tesis. Esto implica repositorios de objetos de aprendizaje, protocolos y estándares de metadatos, ontologías, lenguaje de programación, IDE de desarrollo, etc., de esta manera se facilitaría la elección de la más indicada.

- Diseño: Una vez seleccionada las herramientas más apropiadas se realiza todo el diseño de la aplicación, esto implicará los respectivos diagramas de base de datos, UML, etc., los mismos que permitan definir los distintos parámetros para el desarrollo de la aplicación.
- Implementación: En esta fase se desarrollará la aplicación para la automatización de las anotaciones semánticas.
- Caso de Estudio: Aquí se determinará un caso de estudio para la aplicación desarrollada.
- Pruebas: Esta última fase consistirá en realizar pruebas de la aplicación desarrollada, esto implicará realizar la población de distintas ontologías, consultas sobre dichas ontologías, etc.

1.8. Trabajos Relacionados

1. **Optimizing RDF Data Cubes for Efficient Processing of Analytical Queries:** Este artículo describe contribuciones que se realizan en el estudio de los Cubos de Datos, el primero es la desnormalización de Cubos de Datos. El segundo es proponer el algoritmo de transformación de la red semántica OLAP Denormalizer (SWOD) que convierte un cubo de datos RDF en un cubo o en un patrón completamente desnormalizado. Y el tercero es proporcionar una extensa evaluación experimental, lo que demuestra que los patrones propuestos permiten intercambios efectivos entre el espacio de almacenamiento, tiempos de carga y rendimiento al realizar consultas [12].

2. **Visualizing RDF Data Cubes using the Linked Data Visualization Model:** Este artículo trata sobre la visualización de datos multidimensionales haciendo uso de las herramientas creadas con este propósito. Se parte de un concepto del proceso de visualización de los datos, este proceso consta de cuatro etapas: datos de origen, abstracción analítica, abstracción de visualización y vista. Se menciona que los datos deben pasar por estas etapas dependiendo el visualizador que se vaya a utilizar, ya que los datos deben ser transformados a un formato que sea compatible con este. Los Cubos de Datos facilitan el análisis de los datos ya que se permite que sean visualizados de manera estadística. Dentro de este artículo se menciona el uso de Payola y CubeViz, que son herramientas de visualización amigables para el usuario y su obetivo es presentarle al usuario final información estadística [9].

3. **A Quantitative Survey on the Use of the Cube Vocabulary in the Linked Open Data Cloud:** En este artículo trata sobre el aumento de datos estadísticos en la nube y como el vocabulario de Cubo de Datos se ha convertido en us estándar para el manejo de los datos multidimensionales. Se realiza también un estudio cuantitativo sobre cómo los principales conceptos del vocabulario Cube se aplican en la práctica, utilizando los conjuntos de datos gubernamentales que se han identificado en el LOD 2014 cloud census. El enfóque se realiza más en las estrategias de uso común para el modelado multidimensional

utilizando el Cubo de Datos porque tienen un impacto en la ubicación automática y el consumo de datos [7].

4. **CubeQA—Question Answering on RDF Data Cubes:** En este artículo se trata sobre la influencia del almacenamiento de datos en estructuras multidimensionales, dado que la información se puede manejar estadísticamente y ayuda en la toma de desiciones en áreas como la atención de la salud, las políticas y las finanzas. Lo que se realiza para este artículo es un algoritmo que lo llaman "CubeQA" el cual convierte una pregunta de lenguaje natural en una consulta de SPARQL usando una tubería lineal, en donde primerso se procesa, se encuentran las coincidencias y para finalmente convertir el texto en una consulta SPARQL que se ejecuta para generar el conjunto de resultados que contiene la respuesta, ya que para acceder a los datos dentro del Cubo de Datos en RDF se lo hace mediante consultas utilizando el lenguaje SPARQL [18].

Capítulo 2

Marco Teórico

En este capítulo se describen los fundamentos teóricos y las herramientas que serán utilizadas en el desarrollo de este trabajo de titulación.

2.1. Web Semántica

La web semántica pretende dar un significado a los datos que se encuentran en la web, por lo tanto, este capítulo inicia con una introducción a la web y los cambios que han surgido desde sus inicios, posteriormente se indican ciertos problemas que ha causado la cantidad de información existente y la manera en que la web semántica pretende ayudar.

2.1.1. Introducción

La Web Semántica se define como "Una extensión de la Web actual, en donde la información tiene significado bien definido, para mejorar la cooperación entre computadores y personas" [17]. La Web Semántica puntualiza que no solo las personas deben entender el contenido de la web, sino que también el software, de esta manera se hace posible que haya una comunicación entre personas y ordenadores debido a la semántica que da un significado a la web.

Desde sus inicios la web ha permitido que varias de las tareas que se realizaban manual-

mente sean automatizadas, por ejemplo, el uso de correos electrónicos, la aparición de las redes sociales, y nuevas tecnologías que han surgido en la actualidad, especialmente la tecnología móvil. El avance de la tecnología ha contribuido en el éxito de la web, sin embargo, estos factores que han hecho parte del éxito, también han originado sus principales problemas, como lo es la sobrecarga de información por parte de fuentes no validadas [4]. La Web Semántica ayuda a resolver estos problemas permitiendo a los usuarios establecer tareas dentro de un software que sea capaz de entender textualmente lo que el usuario necesita. El software es capaz de entender lo solicitado , por lo tanto, procesa el contenido que por semántica tiene un significado para el software, permitiéndole razonar y realizar deducciones lógicas que le permitan resolver problemas automáticamente [5] [31].

En la Figura 2.1. se representa una estructura de la web actual versus la web semántica, en donde la web actual interactúa entre sí a través de páginas HTML que contienen información independiente de otra, la web semántica es diferente, la información está representada mediante la estructura de un grafo y no es independiente, al contrario, todos los datos están relacionados haciendo que las búsquedas recorran el grafo devolviendo una respuesta más exacta. Cada grafo esta entrelazado dando un significado a la información que contiene.

El desarrollo de la Web Semántica está enfocado a la construcción de una base de conocimientos sobre las preferencias de los usuarios, por lo tanto, entre la capacidad de conocimiento y la información disponible en Internet, es capaz de atender de forma exacta las demandas de información por parte de los usuarios en relación [22]. Los buscadores actuales lo que hacen es combinar la relación de ciertas palabras con la información disponible en la web, por lo tanto, el resultado no es al 100 % el esperado, por otro lado, la web semántica hace uso de esta ”Base del Conocimiento” y da un significado a los datos permitiendo que el ordenador entienda textualmente la búsqueda ingresada por el usuario, obteniendo así resultados más exactos.

Como se muestra en la Figura 2.2 una Base de Conocimientos está compuesta de cuatro elementos, los cuales se describen como:

- **Agentes Inteligentes:** Estos agentes pueden razonar y aprender a partir de la información que recogen, además de poder tomar decisiones de manera autónoma [19].

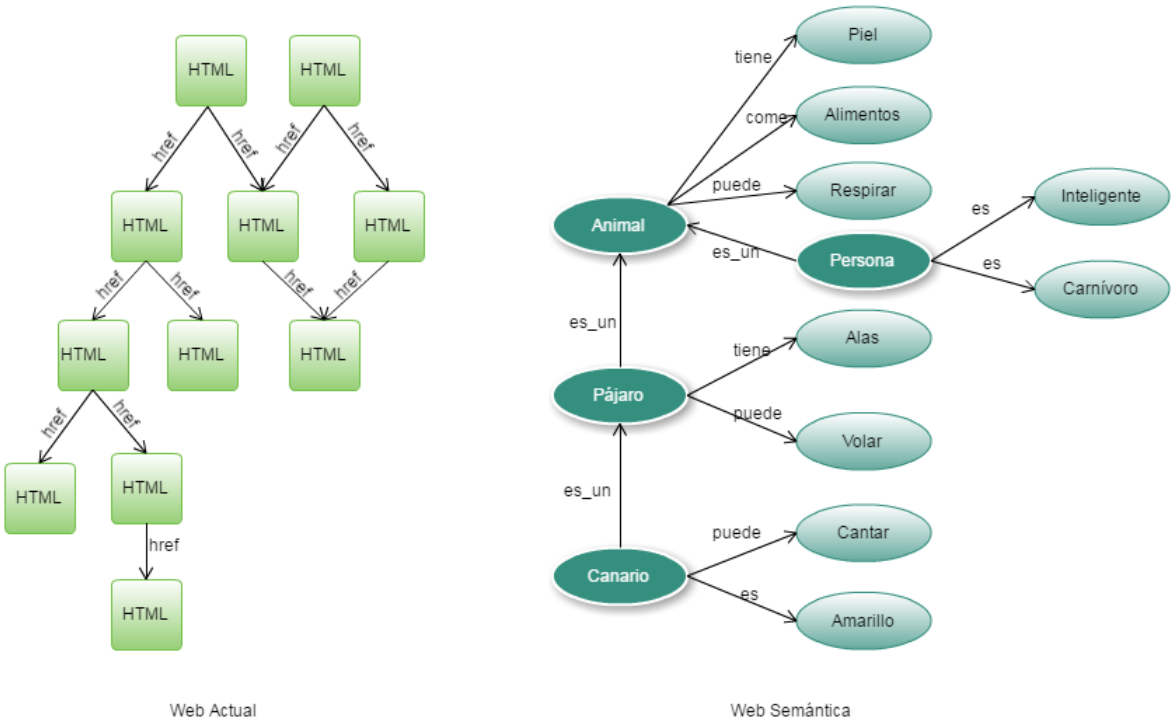

Figura 2.1: Representación de la Web Actual vs. La Web Semántica.

- **Inteligencia Compartida en la Red:** La inteligencia compartida parte de los niveles que identifican la actividad humana, tomando como referente el cúmulo de datos, información y conocimiento, su procesamiento en dirección a la acción, obtenida del ambiente o entorno competitivo [20].

- **Cerebro Glogal:** Los avances en la Web Semántica han sido notables, lo que se pretendía desde sus inicios era el desarrollo de una base de conocimientos que equiparan el funcionamiento de la Web al funcionamiento de un cerebro global[23].

- **Web Semántica:** La Web semántica sería una red de documentos "más inteligentes"que permitan, a su vez, búsquedas más inteligentes [24].

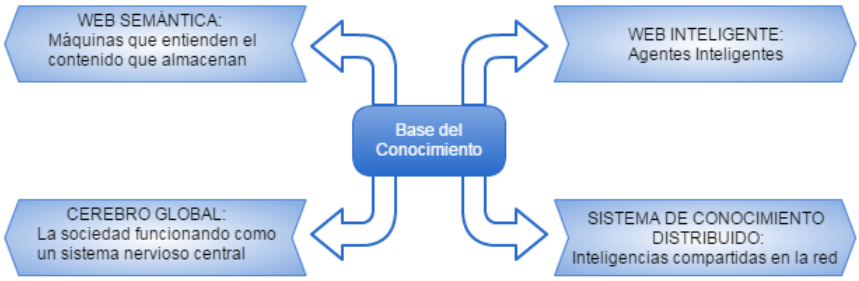

Figura 2.2: Elementos de una Base del Conocimiento.[8]

2.2. RDF

RDF se define como "Un lenguaje descrito que sirve para representar información sobre recursos en la World Wide Web, especialmente propuesto para la representación de Metadatos sobre recursos web. RDF es un modelo de datos que es dispuesto para representar el conocimiento sobre recursos web. Por lo tanto, se basa en la idea de que los objetos a describir se identifican como recursos que poseen propiedades y éstos a su vez tienen valores" [32]. Esta definición se representa de forma gráfica en la Figura 2.3, en donde se identifica un recurso con propiedades y éstos con valores.

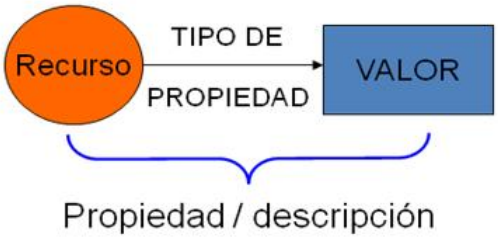

Figura 2.3: Representación de la estructura de un objeto en lenguaje RDF.

Los elementos para una estructura RDF son los siguientes:

- **Recursos:** Los recursos son las cosas descritas por expresiones RDF, los recursos se designan siempre mediante URIs (Uniform Resource Identifier) y la extensibilidad de estos,

permite la introducción de identificadores para cualquier entidad imaginable ya que permite identificar elementos únicos dentro de la web [3].

- **Propiedades:** Las propiedades son un aspecto específico, una característica, atributo o relación que puede utilizarse para describir un recurso. Cada propiedad tiene un significado específico, define sus valores permitidos, los tipos de recursos que puede describir y sus relaciones con las propiedades [3].

- **Declaraciones:** Una declaración RDF es una propiedad más el valor de dicha propiedad para un recurso específico. Una declaración se la conoce también como una sentencia o enunciado y está compuesta por tres partes individuales:

  - **Sujeto:** Recurso
  - **Predicado:** Propiedad
  - **Objeto:** Es el valor de la propiedad que representa el complemento de la misma, éste valor puede también ser otro recurso o puede ser un literal; es decir, un recurso (especificado por una URI) o una cadena simple de caracteres u otros tipos de datos primitivos definidos por XML. [29]

Por ejemplo del enunciado: “Un autor tiene publicaciones” los tres elementos que se distinguen son:

1. El sujeto es: ***autor***
2. La propiedad es: ***tiene***
3. El valor es: ***publicación***

En la Figura 2.4 se identifica la construcción básica de un RDF que es el ”triple” o sentencia, que consiste en dos nodos (sujeto y valor) unidos por un arco (predicado), donde los nodos representan recursos, y los arcos propiedades. Es decir, que el objeto o valor puede a la vez ser el recurso de una siguiente tripleta.

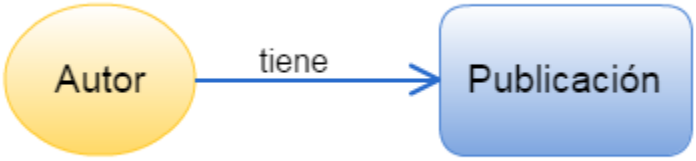

Figura 2.4: Representación de un enunciado en RDF.

El lenguaje RDF es comúnmente utilizado cuando la información necesita ser procesada por aplicaciones que intercambian información que debe ser entendido por ordenadores, más que por humanos. RDF puede utilizarse en diferentes áreas, una de ellas la recuperación de recursos para los buscadores en los cuales mediante el uso de tripletas y las relaciones que tienen entre sí, los resultados al emplear búsquedas llegan a ser casi exactas debido a que la información almacenada posee semántica, es decir que tiene significado propio [3] [32].

2.3. Ontología

Una ontología es una especificación explícita y formal de una conceptualización compartida. Una conceptualización es una vista simplificada y abstracta del mundo que se desea representar para algún propósito en específico, definiendo un vocabulario controlado[6].

Las ontologías definen los términos básicos y las relaciones que comprenden el vocabulario de un tema de alguna área, las reglas para combinar los términos y las relaciones para definir extensiones al vocabulario, es decir, define entidades, clases, propiedades predicados y relaciones entre estos componentes. Debido a que la ontología es un sistema de representación del conocimiento que resulta de la selección de un ámbito o dominio de conocimiento, las ontologías se pueden organizar en estructuras jerárquicas, las cuales se pueden considerar como una de las mejores formas para representar el conocimiento [2].

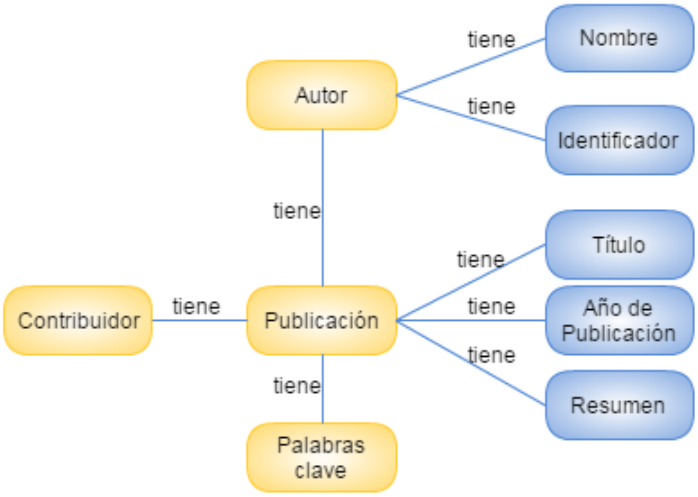

Figura 2.5: Representación de la estructura de un objeto en lenguaje RDF.

2.4. SPARQL

SPARQL es un lenguaje recomendado por la W3C que sirve para realizar consultas sobre datos en RDF, su función es recorrer todas las tripletas dentro del grafo a ser consultado, lo primero que identifica es la concordancia de patrones por cada tripleta, posterior ejecuta ciertos modificadores de solución (de acuerdo a la consulta) y finalmente es la salida en donde realiza la construcción de nuevas tripletas. Los resultados de las consultas SPARQL pueden ser conjuntos de resultados o grafos RDF. SPARQL es el único lenguaje de consultas sobre datos en RDF [25].

2.5. Linked Data

Los ”Datos Enlazados” es la forma que tiene la web semántica de vincular los distintos datos que están distribuidos en la web, estos se referencian de la misma forma que lo hacen los enlaces de las páginas web. La direferencia entre la realción de las conexiones de la web actual, donde los enlaces son relaciones entre puntos de los documentos escritos en HTML, los datos enlazan cosas arbitrarias que se describen en RDF. El objetivo central es que por medio de Linked Data se pueda construir en Internet una gran base de datos en donde se pueda llevar a cabo consultas especificas. [30]

2.6. Apache Marmotta

Apache Marmotta proporciona una especificación sobre el uso de HTTP para interactuar con los servidores que exponen sus recursos como Linked Data. El objetivo es proporcionar una implementación abierta de una plataforma de Linked Data que puede ser utilizado, extendido y fácilmente desplegado por las organizaciones que desean publicar los datos vinculados o construir aplicaciones personalizadas en Linked Data. [1]

2.7. Data Warehouse

Un Datawarehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde varias perspectivas y con grandes velocidades de respuesta. Esta base constituye un conjunto de datos integrados y orientados a elementos concretos y generales, que varían con el tiempo y que no son transitorios, además soportan el proceso de ayuda en la toma de decisiones de la administración [16].

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información. Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma [11].

Según definió Bill Inmon[21], el Data Warehouse se caracteriza por ser:

- **Integrado:** Los datos almacenados en el Data Warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.
- **Temático:** Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

- **Histórico:** El tiempo es parte implícita de la información contenida en un Data Warehouse. La información almacenada en el Data Warehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.
- **No volátil:** El almacén de información de un Data Warehouse existe para ser leído, y no modificado. La información es por tanto permanente, significando la actualización del Data Warehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

Figura 2.6: Elementos que componen un Data Warehouse.

La Figura 2.6 indica gráficamente la definición de un Data Warehouse, el cual se divide en tres partes: la primera muestra las diversas fuentes de información que puede ser cualquier contenedor de datos, la segunda define el proceso de Data Warehouse necesario para integrar los datos en una Base de Datos unificada y la tercera sección especifica los resultados, siendo estos de gran interés debido a que sirve y ayuda a las organizaciones en la toma de decisiones debido a que mantiene la información consolidada.

2.8. Cubo de Datos en RDF

Con el incremento masivo de datos dentro de la web en la actualidad, surge la necesidad de llevar los modelos de datos mencionados en el punto anterior, a modelos en RDF, por lo que se

debe buscar alternativas para mejorar las búsquedas dentro de la web. El almacenamiento de datos multidimensionales es una de las alternativas para dar solución a este problema, además de que los datos multidimensionales están vinculados con la representación de resultados estadísticos. Para el almacenamiento y el manejo de datos multidimensionales la W3C recomienda el uso del **Vocabulario del Cubo de Datos**, éste vocabulario se centra exclusivamente en la publicación de datos multidimensionales en la web. Un ejemplo claro de los datos necesarios a ser publicados son los datos estadísticos, estos datos crean una base para la predicción de políticas, la planificación y ajustes, y sustenta muchas de las combinaciones de datos y visualizaciones que se presentan en la web [26].

El Cubo de Datos se basa en los siguientes vocabularios RDF existentes [26]:

- **SKOS** para esquemas de conceptos.
- **SCOVO** para estructuras estadísticos básicos.
- **Términos de Dublin Core** para metadatos.
- **VoiD** para el acceso a datos.
- **FOAF** para los agentes.
- **ORG** para organizaciones.

Cuando los datos(especialmente datos estadísticos) estan siendo publicados en la web, este vocabulario permite darle un significado semántico mediante la utilización del estándar RDF. Es así que el modelo del Cubo de Datos esta compuesto por clases, propiedades, observaciones y operaciones que pueden ser aplicadas a un conjunto de datos, estos son componentes del vocabulario del Cubo de Datos, cuyo objetivo es tener la información de manera multidimensional.

El vocabulario facilita la representación de un conjunto de datos estadísticos que comprende de una colección de observaciones hechas en algunos puntos a través de la relación del mundo real y la forma con la que es representado gráficamente en el vacabulario. La colección puede ser caracterizada por un conjunto de dimensiones que definen la observación (por ejemplo, tiempo, área, género), un conjunto con metadatos que describen lo que se ha medido (por ejemplo,

la actividad económica, la población), la forma en que se midió y cómo las observaciones se han expresado (por ejemplo, unidades, multiplicadores, variables de estado). Se puede pensar en los datos estadísticos establecidos como un espacio multidimensional, o hiper-cubo, indexada por esas dimensiones. Las dimensiones que se decidan dar al Cubo de Datos no dependen de un número determinado, caso contrario, el número de dimensiones puede ser un número especificado por el desarrollador [27].

La Figura 2.7 muestra el esquema del vocabulario propuesto por la W3C, que especifica la estructura para definir Cubos de Datos y almacenar datos en estructuras dimensionales.

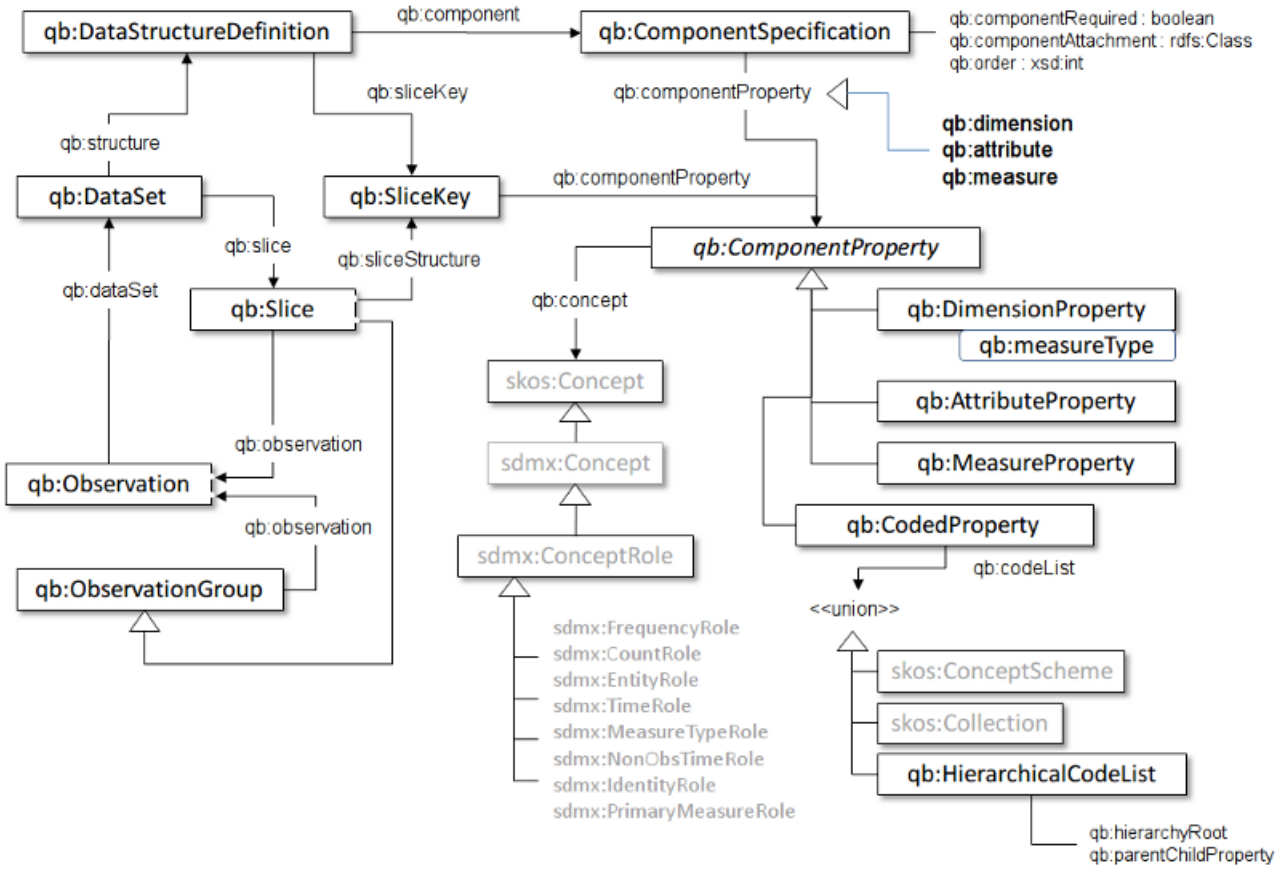

Figura 2.7: Vocabulario RDF Data Cube.

Como se mencionó anteriormente, la estructura del vocabulario del Cubo de Datos está compuesto por clases (representado en la Figura 2.7 por rectángulos) y propiedades (representado en la Figura 2.7 por las líneas que unen los rectángulos), las clases representan los elemen-

tos que componen el Cubo de Datos y las propiedades definen el tipo de relación que existe entre las clases .

La representación gráfica del vocabulario del Cubo de Datos inicia con la definición de la estructura del Cubo de Datos y a partir de éste se dividen en componentes que son específicos para formar el Cubo de Datos. La entidad **qb:DataStructureDefinition** define la estructura de uno o mas conjuntos de datos, la entidad **qb:DataSet** representa al conjunto de datos, **qb:Observation** es el valor de la medida, **qb:ComponentProperty** esta propiedad define las subclases que se detallan a continuación:

- **qb:DimensionProperty**, representan las dimensiones dentro del Cubo de Datos.
- **qb:AttributeProperty**, esta propiedad permite calificar e interpretar el valor observado, además permite obtener la especificación de las unidades de medida.
- **qb:MeasureProperty**, representa al valor de la medida para el fenómeno observado.

Además indica que el Cubo puede ser dividido en Slice utilizando la propiedad **qb:Slice** que representa un subconjunto del conjunto de datos que se forman a partir de un subconjunto de valores dimensionales.

Almacenando los datos en una estructura multidimensional se asegura que las consultas hacia estos datos sean mas rápidas, esto se da debido a que se accede unicamente a las dimensiones solicitadas y no necesita consultar los datos de todas las dimensiones en el Cubo de Datos.

2.9. Ejemplo de transformación de RDF a RDF Data Cube

Partiendo de lo difícil que sería acceder a los datos si estos estuvieran desorgianizados y si no existiera una forma sistemática para poder recuperar esos datos, se puede decir que las bases de datos juegan un papel importante con el usuario al momento de interactuar con el almacenamiento y recuperación de información. Actualmente existen varias alternativas para realizar estos procesos, en este ejemplo los datos inicialmente se encuentran en estructuras

similares a un grafo, bajo el estándar de RDF y lo que se pretende es cambiar la estructura de los datos hacia una estructura multidimensional.

Inicialmente se mencionó que un RDF está compuesto de entidades, propiedades y relaciones, estos elementos se pueden observar gráficamente mediante un grafo que es la representación de un texto al que se le da un significado. Por ejemplo, en la Figura 2.8, se tiene al autor "Víctor Saquicela" que tiene publicaciones y esta a su vez tiene palabras clave, contribuidores y pertenece a una fuente en donde está almacenada la información. El grafo indica como está estructurado un RDF para que permita realizar búsquedas.

Para el ejemplo se utilizarán los datos que son provenientes del proyecto REDI, que se detallará en los siguientes apartados, éstos datos están en formato RDF y actualmente se encuentran unificados en un grafo central.

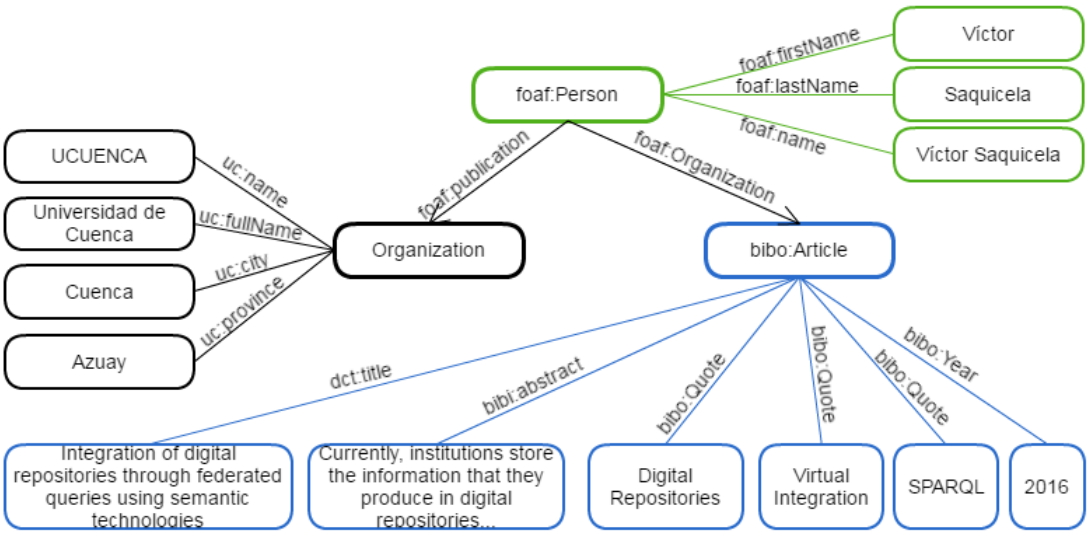

Figura 2.8: Representación gráfica en RDF.

La información representada en la Figura 2.8 muestra la organización de los datos en RDF, a partir de esta estructura el objetivo es obtener una estructura multidimensional, partiendo del vocabulario del Cubo de Datos se puede identificar las entidades o clases y las subclases. En la siguiente tabla se puede observar los nodos seleccionados dentro del grafo central que contiene los datos en RDF y además las entidades dentro del vocabulario del Cubo de Datos con las que se relacionan.

NODOS EN EL GRAFO RDF	ENTIDADES DEL VOCABULARIO DEL CUBO DE DATOS
Conjunto de datos	qb:DataSet
Publicación	qb:MeasureProperty qb:AttributeProperty
Autor	qb:DimensionProperty
Coautor	qb:DimensionProperty
Palabra(s) clave	qb:DimensionProperty
Fuente	qb:DimensionProperty
Año de Publicación	qb:DimensionProperty

Cuadro 2.1: Relación entre nodos y el vcabulario del Cubo de Datos

Para identificar los nodos y las propiedades del vacabulario del Cubo de Datos, primeramente se debe establecer el elemento sobre el cual se va a almacenar la información, para el ejemplo, el elemento principal es la publicación, es decir que entorno a las publicaciones se va a obtener y almacenar en la estructura multidimensional la información. Por los tanto, la publicación es la medida dentro del vocabulario del Cubo de Datos y a partir de ella se obtiene las dimensiones que serán los que complementan la información de cada publicación.

Por lo tanto, las publicaciones serán los valores a ser observados, en el vocabulario de Cubo de Datos está representado por la entidad **qb:Observation**, a partir de la publicación se derivan otros componentes como son: autores, coautores, palabras clave, y fuente de obtención de datos. Éstos componentes en el Cubo de Datos son vistos como dimensiones y están representados por la entidad **qb:DimensionProperty**.

Ya que los valores han sido identificados, en la siguiente tabla se muestra una tabla multidimensional en donde se representa de manera general como se organizarían los datos que inicialmente estaban con la estructura de una grafo como se mostró en la Figura 2.8. También se puede observar que las dimensiones son independientes entre sí, lo que se podría decir que ésta es una de las ventajas de la estructura multidimensional, debido a que solamente accede a las dimensiones solicitadas por el usuario, liberando de esta manera sobrecarga de información no necesaria en determinadas consultas de datos. Por ejemplo, si el usuario realiza una búsqueda para obtener el número de publicaciones de un determinado autor, el buscador realizará la búsqueda en la dimensión de los autores y la medida, que es la publicación, para determinar el

número de incidencias y de esta manera obtener el resutado esperado.

		2014-2016		
		Google Scholar		
		José Segarra	Mauricio Espinoza	José Ortiz
Víctor Saquicela	Linked Data	Integration of digital repositories through federated queries using semantic technologies		
	Digital Repositories			
	Virtual Integration			
	SPARQL			

Cuadro 2.2: Tabla de Datos Multidimensionales.

Los datos que se encuentran en la Tabla 2.2 son reorganizados a un Cubo en tres dimensiones en donde se puede identificar como estaría almacenada la información dentro del Cubo de Datos. Un Cubo de Datos puede contener de uno o mas cubos dentro del mismo, así como se muestra en la Figura 2.9

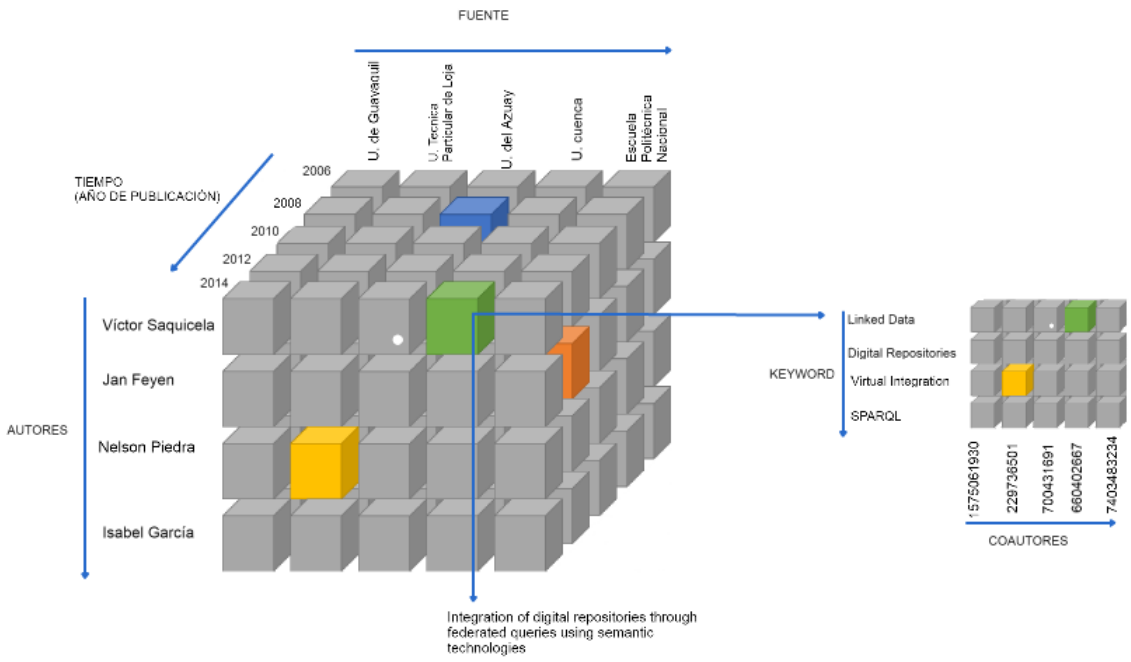

Figura 2.9: Representación gráfica del Cubo de Datos.

Es así entonces como se representa gráficamente el almacenamiento de los datos dentro de

la estructura multidimensional, haciendo uso del vocabulario del Cubo de Datos.

2.10. Open Cube Toolkit

OpenCube Toolkit es una herramienta de código abierto y está disponible en su página oficial ”http://opencube-toolkit.eu/”, OpenCube Toolkit está diseñada específicamente para visualizar información multidimensional de forma gráfica.

El proyecto OpenCube en general y sus componentes , en particular, se centran en el procesamiento de cubos de datos RDF: datos multidimensionales representados como RDF y estructurados de acuerdo a la ontología RDF Data Cube. La mayoría de los componentes desarrollados se dirige a la etapa de utilización de datos para su visualización estadística y más orientado a los usuarios finales en lugar de administradores de datos, además permite crear aplicaciones personalizadas y la realización de las funciones de bajo nivel genéricos tales como el acceso de datos compartidos, registro y seguimiento [10].

2.10.1. El ciclo de vida de OpenCube

El ciclo de vida de OpenCube inicia con la obtención de datos que han sido procesados y manipulados previamente, a fin de obtener datos dimensionales con los que posteriormente se obtiene información gráfica. La representación gráfica del ciclo de vida de OpenCube se muestra en la Figura 2.10

Las etapas del ciclo de vida se clasifican en dos grandes fases (a) la fase que incluye la definición y estructura de un cubos de datos, y (b) la fase de reutilización que incluye el aprovechamiento ligado de cubos de datos en análisis avanzados y publicación de las visualizaciones [15].

Ciertas herramientas que apoyan todas las etapas del ciclo de vida correspondiente para el Cubo de Datos son:

- **OpenCube Browser** que permite la exploración de un cubo de datos vinculado con la presentación por slice (secciones dentro del cubo de datos) de dos dimensiones del cubo en

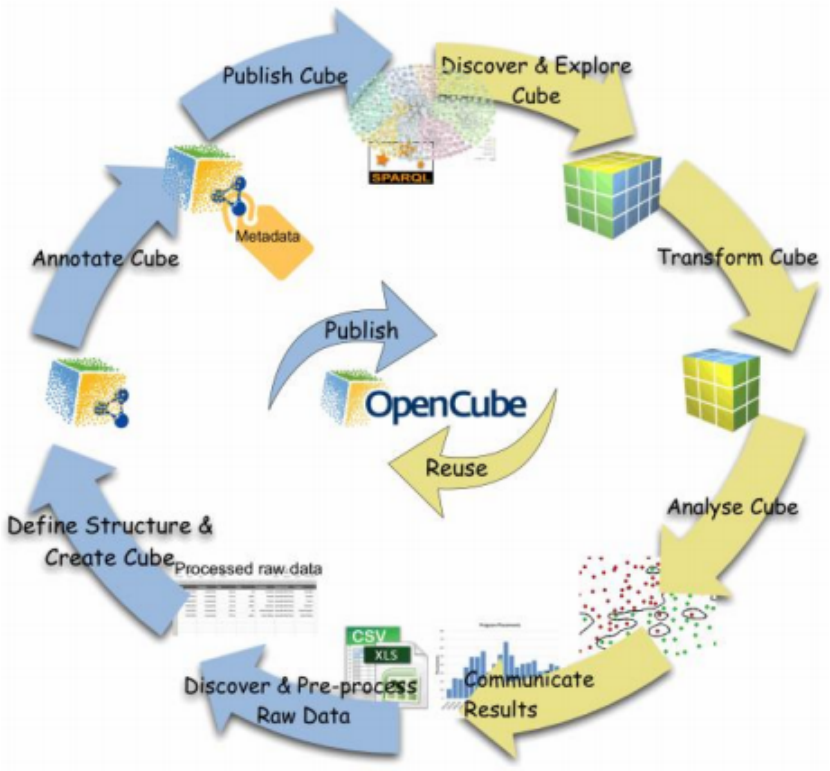

Figura 2.10: Ciclo de vida de OpenCube [15].

forma de tabla. Esta herramienta también es compatible con las operaciones OLAP.

- **OpenCube MapView** que permite la visualización de cubos de datos vinculados en un mapa basado en su dimensión geoespacial. En la actualidad, la MapView soporta marcadores, burbujas y mapas choropleth. También es compatible con las operaciones OLAP como para permitir la visualización de varias vistas de un cubo.

- **Herramientas interactivas de visualización gráfica** que permite vincular cubos de datos con un resumen de las visualizaciones de datos y la creación de gráficos.

Algunas herramientas de presentación gráfica de información son los siguientes:

- **TreeResult:** Permite la visualización jerárquica del Cubo de Datos en un árbol definiendo sus funcionalidades principales que son la especificación de una query que se requiere visualizar.

- **TableResult:** Permite la visualización y configuración de tablas, se puede definir una query para determinar los datos que serán visualizados, el título, etiquetas, configuración de columnas, mensajes y dimensiones.

2.11. REDI - Repositorio Semántico de Investigadores del Ecuador

Es un proyecto realizado conjuntamente entre CEDIA y el Departamento de Ciencias de la Computación - U. de Cuenca. Este proyecto de investigación se centra en la detección automática de áreas similares de conocimiento entre los investigadores del Ecuador. Su objetivo es apoyar y fortalecer las estrategias de búsqueda de datos acerca de investigadores, manteniendo un repositorio común de datos. Para ello, se define e implementa una arquitectura de software que permite realizar una búsqueda eficiente de investigadores acerca de sus publicaciones que estarán al servicio de las universidades miembros de CEDIA, además que permita detectar áreas similares de conocimiento [14].

Este proyecto surgió con el objetivo de identificar áreas similares de investigación, ya que recoge la información de todos los investigadores ecuatorianos con sus respectivas publicaciones a fin de agrupar áreas similares de conocimiento.

Inicialmente se han identificado los repositorios de publicaciones en donde está almacenada la información, estos son: Scopus, Microsoft Academics, Google Scholar, etc. [14]. Por ejemplo, si hay una publicación de un determinado autor y se encuentra dentro de estos repositorios, se puede deducir que la información relacionada a dicha publicación, no es exactamente lo mismo lo que contiene un repositorio a lo que contiene otro repositorio. Por la tanto, se puede decir que la información puede o no estar completa dentro de cada repositorio, pero como sabe el usuario que la información percibida es completa. REDI trata de unificar toda la información dentro de un grafo central que obtenga la información de todos estos repositorios de publicaciones, a fin de tener un solo almacenamiento de datos y el usuario de esta manera obtendrá información recolectada de todos los RDP (Repositorios de Publicaciones) y no tendrá necesi-

dad de consultar en todos los buscadores por separado , ya que obtendrá toda la información dentro de la misma herramienta.

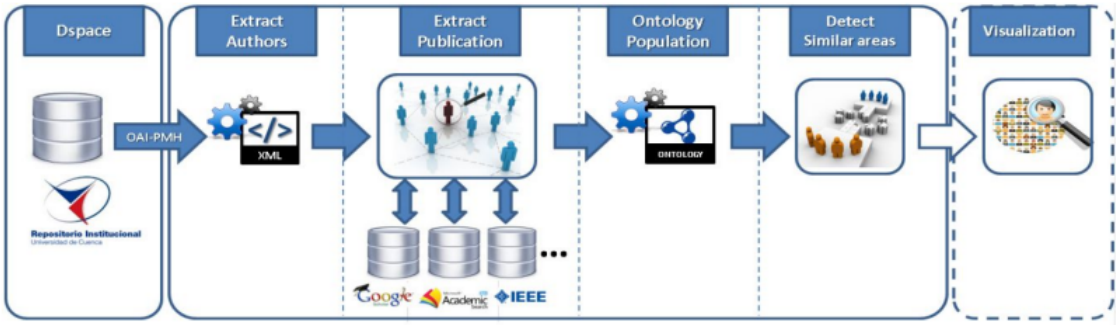

Figura 2.11: Arquitectura del REDI. [14]

La Figura 2.11 muestra la arquitectura general del REDI, en donde se distinguen cinco procesos:

1. **Extracción de autores:** Consiste en extraer datos de autores que se encuentran en repositorios digitales, éstos datos son transformados a RDF y almacanados en un modelo ontológico.

2. **Extracción de publicaciones:** Consiste en extraee datos de publicaciones que se encuentren en los RDP, estas publicaciones corresponden a los autores que se han recuperado en el proceso anterior, una vez extraída la información esta es almacenada en un modelo ontológico.

3. **Población de la ontología:** La información extraida de los autores y las publicaciones se almacenan en un grafo Triple Store (base de datos para almacenamiento y recuperación de tripletas en rdf). La información es intregrada y consolidada en un grafo central.

4. **Detección de áreas similares de conocimiento:** Se realizan ciertas técnicas para detectar patrones que permitan identificar similitudes en la información almacenada en el grafo central para finalmente esta información ser anotada semánticamente en el modelo ontológico.

5. **Visualización:** Para la visualización se ha implementado una aplicación web que permite a los usuarios realizar búsquedas de publicaciones de diferentes autores, la información es visualizada utilizando modelos de visualización que permiten navegar dentro del grafo, haciendo que el usuario final tenga un fácil e intuitivo acceso a la información disponible. En la Figura 2.12 se puede observar como se puede obtener un autor con todas sus publicaciones y los contrubuidores de la misma, al mismo tiempo si se espera ver las publicaciones de los contribuidores basta con ingresar al grafo y se despliega una nueva lista de sus publicaciones.

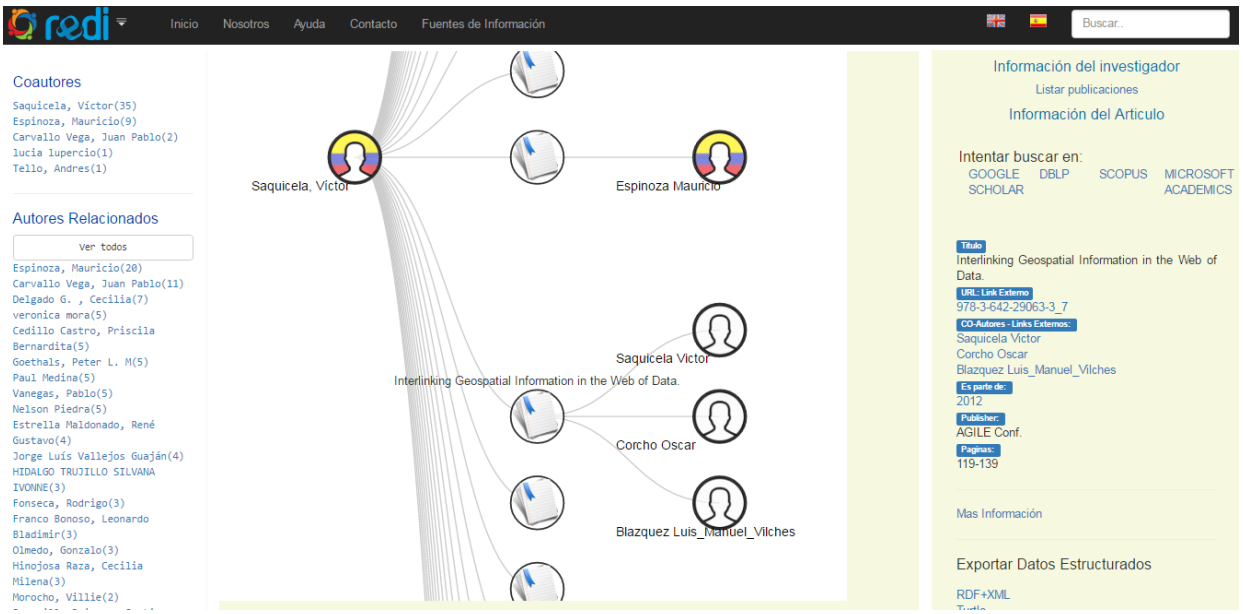

Figura 2.12: Visualización de autor "Víctor Saquicela", publicaciones y contribuidores.

En el presente trabajo de titulación se utilizará un nuevo modelo ontológico basado en el Cubo de Datos, lo que permitiría reorganizar los datos que actualmente se encuentran en grafos en RDF hacia datos dimensionales que permitan una nueva estructura en donde se espera que los tiempos de acceso a los datos al momento de realizar las búsquedas sean más cortos, además de utilizar un visualizador propio para datos dimensionales, el mismo que será incorporado dentro del sistema del REDI ofreciendo nuenvas alternativas de búsqueda a los usuarios finales.

Capítulo 3

Proceso de obtención y mapeo de datos

En este capítulo se describe el desarrollo práctico realizado para obtener los datos con los que se trabajará a lo largo del proceso de transformación de datos, además se mencionan los criterios tomados en cuenta para la definición del modelo del Cubo de Datos.

3.1. Introducción

Como se mencionó en el capítulo 2, el presente trabajo de titulación parte del proyecto REDI. Actualmente, los datos que se están manejando dentro del sistema del REDI están bajo el éstandar de RDF, son datos semánticos, por lo que el objetivo es cambiar la estructura del almacenamiento de datos, hacia una estructura multidimensional, para optimizar búsquedas en el desarrollo del sistema.

Para llevar a cabo el objetivo planteado, inicialmente se desarrolla un modelo gráfico que representa la arquitectura del proceso de transformación de datos, ésta arquitectura muestra un nivel general ya que permite visualizar cada una de las etapas por las que deben pasar los datos antes de obtener una estructura multidimensional semántica para su almacenamiento. Este modelo de arquitectura consta de cuatro etapas, la primera es la obtención de los datos con los que actualmente trabaja el sistema del REDI, dentro de esta etapa además se realiza un mapeo de datos para identificar la organización de los datos dentro del Cubo de Datos, posteriormente pasa por el proceso de transformación de datos, obteniendo así los datos almacenados en la

estructura multidimensional que por medio de determinadas herramientas serán visializados al usuario final. La Figura 3.1 muestra el modelo de arquitectura mencionado.

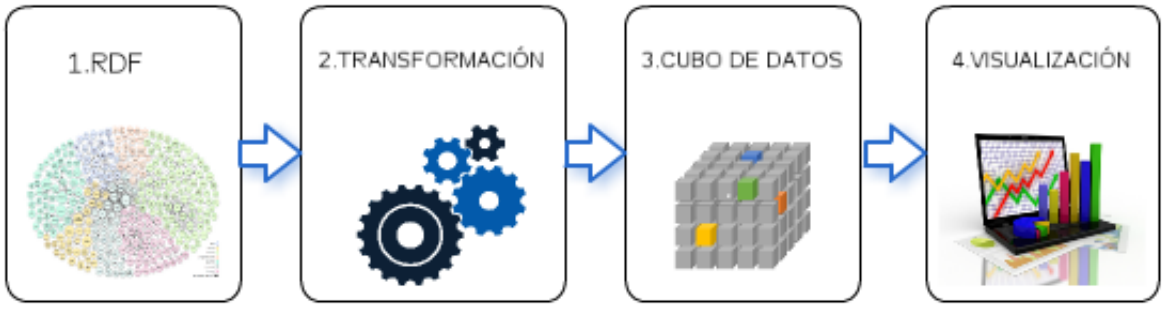

Figura 3.1: Arquitectura de transformación de datos semánticos a datos multidimensionales

En este capítulo se describe la primera etapa de la arquitectura del proceso de trasformación de datos semánticos en RDF hacia datos semánticos multidimensionales, la descripción del proceso realizado para cumplir las siguientes etapas se menciona en los siguientes capítulos.

3.2. Descripción de la ontología RDF utilizada por el REDI

El modelo antológico utilizado para almacenar los datos semánticamente está basado en la ontolgía BIBO (es una ontología para la Web semántica para describir cosas bibliográficas como libros o revistas), y FOAF (es una ontología que describe a las personas, sus actividades y sus relaciones con otras personas y objetos) [13], éstas ontológias están construidas sobre un LDP (Linked Data Platform). Un LDP es una especificación de datos enlazados que define un conjunto de patrones de integración para crear servicios HTTP capaces de leer y escribir datos RDF [28] , el modelo ontológico utilizado por el REDI se muestra en la Figura 3.2.

El modelo de la Figura 3.2 muestra como están anotados semánticamente los datos dentro de la ontología, es decir como están escritos los datos siguiendo la estructura del formato RDF, el REDI utiliza un Triplestore para el almacenamiento y recuperación de tripletas. Un Triplestore es una base de datos para almacenamiento y recuperación de datos en formato RDF mediante consultas SPARQL.

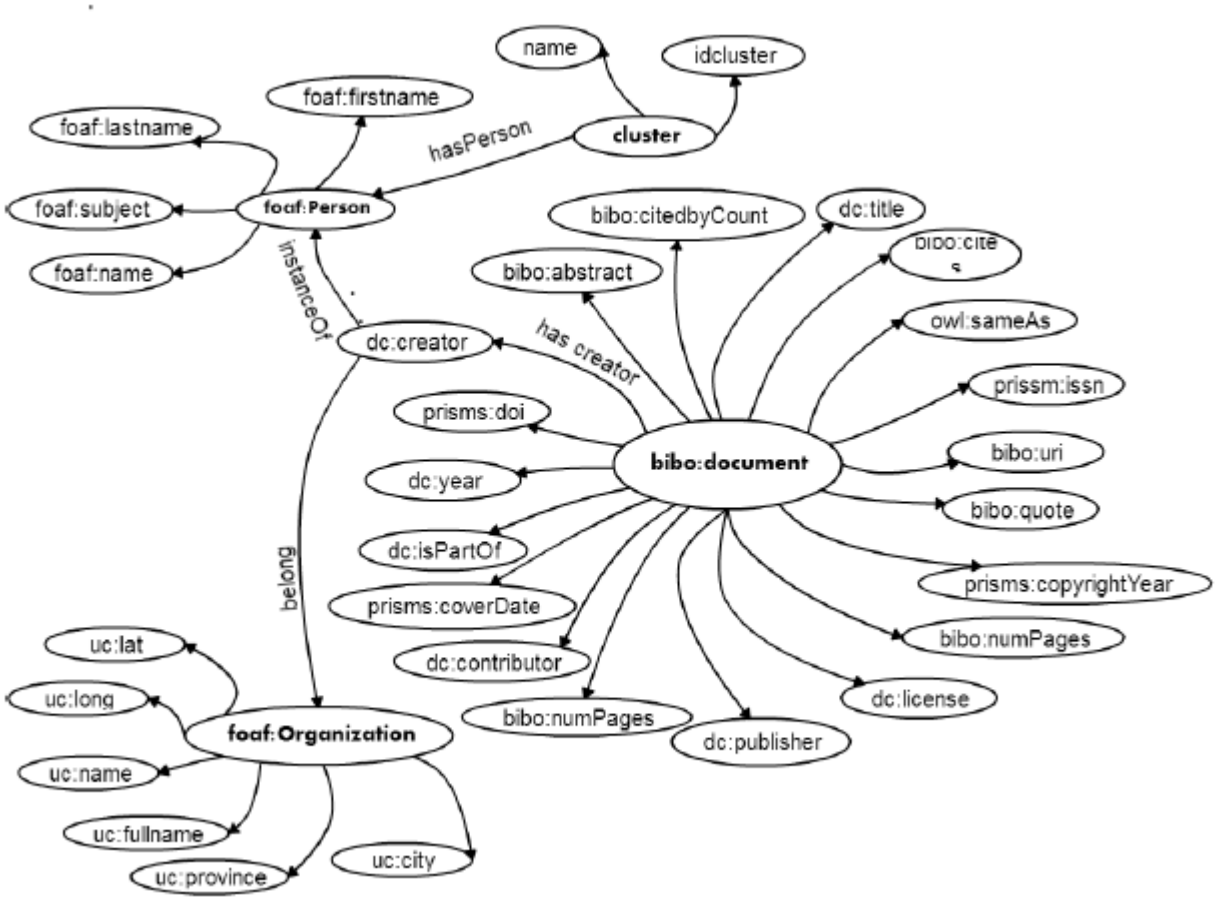

Figura 3.2: Ontología RDF utilizada para el sistema del REDI [13].

Por lo tanto, para almacenar o recuperar la información dentro del Triplestore que contiene todo la información utlizada por el sistema del REDI, basta con realizar consultas SPARQL que sirven para el manejo de datos dentro del grafo central. A continuación se detalla la primera etapa para iniciar el proceso dentro de la arquitectura mostrada en la Figura 3.1.

3.3. Obtención de datos semánticos en RDF

Como ya se mencionó anteriormente, los datos iniciales son obtenidos del sistema REDI que almacena los datos dentro de un grafo central. El lenguaje de consultas utilizado para recorrer las tripletas dentro de la ontología es el lenguaje SPARQL. Por lo tanto para acceder a los datos dentro de la plataforma y obtener una visualización previa de los datos requeridos, se

pueden realizar consultas ingresando al servidor utilizado por REDI mediante la dirección web: "http://redi.cedia.org.ec", así se obtiene los datos necesarios para el proceso de transformación.

El proyecto REDI tiene por objetivo realizar búsquedas relacionando áreas similares de conocimiento, para ello es necesario que las incidencias se realicen sobre las publicaciones dentro del repositorio semántico, por lo tanto, realizando la consulta que se muestra en la Figura 3.3, se obtiene un grupo de tripletas de información relacionada a cada publicación, estos valores son de: autores, coautores, palabras clave, año de publicación y el nombre de la publicación correspondiente.

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX uc: <http://ucuenca.edu.ec/wkhuska/resource/>
PREFIX year: <http://prismstandard.org/namespaces/basic/2.0/>
    SELECT ?UriAutor ?UriPublicacion ?UriContribuidor ?UriFuente ?Keyword ?Anio
         WHERE {
            graph <http://ucuenca.edu.ec/wkhuska>
            {
                ?UriAutor foaf:publications ?UriPublicacion.
                ?UriPublicacion dct:contributor ?UriContribuidor.
                ?UriPublicacion bibo:Quote ?Keyword.
                ?UriPublicacion year:publicationYear ?Anio.
                ?UriAutor dct:provenance ?UriFuente.
            }
        }
```

Figura 3.3: Consulta SPARQL: Información de publicaciones.

EL resultado proveniente de la consulta consta de aproximadamente 50.000 tripletas, las que contienen información extraída de las fuentes de datos (Google Scholar, Scopus, DBLP, etc), es decir que la información obtenida es completa de acuerdo a la información relacionada a cada publicación.

Los repositorios del REDI albergan varios datos con respecto a cada publicación, por lo que se han extraído únicamente seis de los datos que se han considerado necesario para relacionar y juntar la información correspondiente de cada publicación. Los datos que se han extraídos desde el grafo central del REDI, son los datos que serán utilizados para realizar la población de datos hacia una estructura multidimensional semántica. En la Tabla 3.1 se muestra el resultado obtenido al realizar la consulta que se muestra en la Figura 3.3.

Uri Autor	Uri Publicación	Uri Contribuidor	Uri Fuente	Keyword	Anio
http://ucuenca.edu.ec/resource/author/eduardo-gustavo-valarezo-armijos	http://ucuenca.edu.ec/wkhuska/publication/chemical-composition-of-essential-oils-of-two-species-of-the-lamiaceae-scutellaria-volubilis-and-lepechinia-paniculata-from-loja	http://ucuenca.edu.ec/resource/author/eduardo-gustavo-valarezo-armijos	http://ucuenca.edu.ec/wkhuska/endpoint/4d0ebfe0bc494647139f10dfe308551f	Aromadendrene xsd:string	2012 xsd:string
http://ucuenca.edu.ec/resource/author/eduardo-gustavo-valarezo-armijos	http://ucuenca.edu.ec/wkhuska/publication/chemical-composition-of-essential-oils-of-two-species-of-the-lamiaceae-scutellaria-volubilis-and-lepechinia-paniculata-from-loja	http://api.elsevier.com/content/author/author_id/41262539300	http://ucuenca.edu.ec/wkhuska/endpoint/4d0ebfe0bc494647139f10dfe308551f	Aromadendrene xsd:string	2012 xsd:string
http://ucuenca.edu.ec/resource/author/eduardo-gustavo-valarezo-armijos	http://ucuenca.edu.ec/wkhuska/publication/chemical-composition-of-essential-oils-of-two-species-of-the-lamiaceae-scutellaria-volubilis-and-lepechinia-paniculata-from-loja	http://api.elsevier.com/content/author/author_id/6504407012	http://ucuenca.edu.ec/wkhuska/endpoint/4d0ebfe0bc494647139f10dfe308551f	Aromadendrene xsd:string	2012 xsd:string
http://ucuenca.edu.ec/resource/author/eduardo-gustavo-valarezo-armijos	http://ucuenca.edu.ec/wkhuska/publication/chemical-composition-of-essential-oils-of-two-species-of-the-lamiaceae-scutellaria-volubilis-and-lepechinia-paniculata-from-loja	http://api.elsevier.com/content/author/author_id/49964019600	http://ucuenca.edu.ec/wkhuska/endpoint/4d0ebfe0bc494647139f10dfe308551f	Aromadendrene xsd:string	2012 xsd:string

Cuadro 3.1: Información obtenida de la consulta de la Figura 3.3

Como se puede apreciar en la tabla 3.1, la información extraída es:

- **Publicaciones:** Contiene el nombre de la publicación que se ha encontrado en las fuentes de datos.
- **Autores:** Muestra el nombre del autor principal de la publicación.
- **Coautores:** Muestra el nombre de los contribuidores relacionados a la publicación.
- **Palabras clave:** Indica todas las palabras claves incluidas dentro del abstract de la publicación.
- **Año de publicación:** Indica el año de publicación.
- **Fuente:** Indica el nombre de la fuente de donde fue extraída la información.

3.4. Mapeo de Datos RDF de REDI a un modelo de Cubo de Datos

Un mapeo se realiza con el objetivo de mostrar como se corresponden los objetos del modelo inicial con la base destino de los datos. Es decir como se asosian entre si, esta asociación es definida por el desarrollador quien debe describir los elementos relacionados entre las estructuras de almacenamientos de datos. Una vez mencionado el objetivo y la importancia de realizar un mapeo previo al momento de realizar un cambio de estructura sobre las fuentes de almacenamiento de datos, se procede con el mapeo entre la ontología utilizada por el sistema del REDI y lo ontología representada por el vocabulario del Cubo de Datos.
De los datos obtenidos del REDI, el objetivo central sobre el cual se espera tener información son las publicaciones, cada publicación tiene datos que complementan la información correspondiente a cada una, así como se muestra en la Figura 3.4 en donde se describe como se relacionan los datos cuando se organizan dentro de una estructura multidimensional.

Figura 3.4: Relación de los datos en una estructura multidimensional.

La Figura 3.4 indica gráficamente que los datos de cada elemento son independientes, relacionados únicamente por la publicación a la que pertenecen. Es por esta razón que cuando se realizan consultas para acceder a los datos, éstas únicamente acceden a los elementos descritos en la consulta. Por ejemplo, si en la consulta se espera obtener el número de publicaciones con los autores correspondientes, solamente se accede al elemento de los autores y publicaciones, optimizando así los tiempos de respuesta para obtener los resultados de dicha consulta.

La W3C define un vocabulario estándar (ver Figura 2.7) para formar un Cubo de Datos. Con los datos disponibles en la plataforma del REDI, éstos se relacionan al vocabulario mencionado y se obtienen las dimensiones, atributos y medidas que serán empleados para formar el Cubo de Datos. La Tabla 3.2 muestra el mapeo que se realizó para identificar los valores que serán designados como dimensiones, al valor de la medida por defecto se asigna la publicación, debido a que el proyecto REDI tiene por objetivo mostrar todas las publicaciones que están dentro del repositorio web, por lo tanto, se toma este valor como el valor a ser observado, es decir, que las dimensiones que se han identificado serán relacionadas entre sí en función de la publicación a la que pertenecen.

ONTOLOGÍA DEL REDI	**ONTOLOGÍA DEL CUBO DE DATOS**
bibo:uri	qb:DataStructureDefinition
dc:isPartOf	qb:ComponentSpecification
dc:license	____
rdf:type	qb:ComponentProperty
dc:publisher	qb:ObservationGroup
dc:contributor	qb:DimensionProperty
dc:title	qb:Observation
bibo:numPages	____

Cuadro 3.2: Mapeo del Cubo de Datos

Los datos deben ser mapeados para identificar las clases que serán necesarias para integrarlos a la ontología del Cubo de Datos. Como se aprecia en la Figura 3.4, la publicación es el elemento central, es decir, el valor que debe ser observado, y de acuerdo al vocabulario del Cubo de Datos utiliza la instancia **qb:Observation**, por otro lado los elementos restantes son las dimensiones que representan el Cubo de Datos y utilizan la instancia **qb:DimensionProperty**, las demás instancias encontradas en el mapeo de datos representan el conjnto de datos de manera general.

En la tabla 3.2 se identifica que no todas las instancias del vocabulario del Cubo de Datos se relacionan con la ontología semántica en RDF utilizada por el REDI, en consecuencia se obtiene un conjunto de instancias del vacabulario que han sido identificadas en el mapeo, las instancias que serán utilizadas para la anotación semántica dentro del modelo de Cubo de Datos se

presenta en la Figura 3.5.

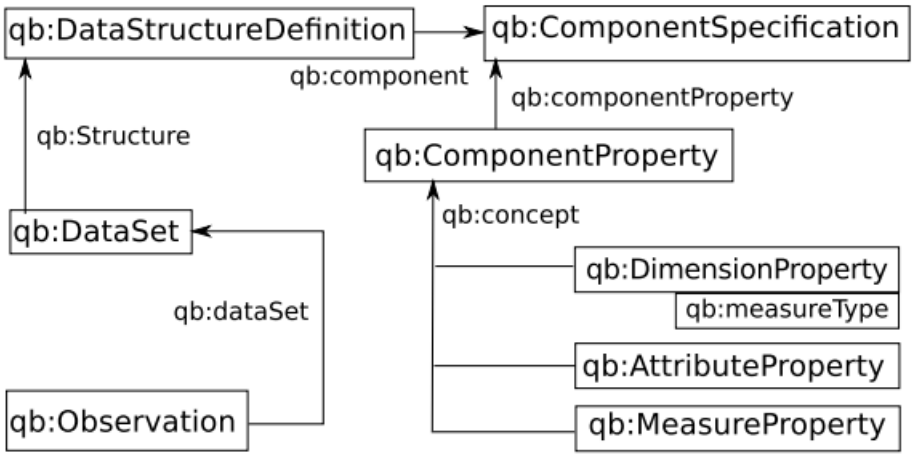

Figura 3.5: Vocabulario del Cubo de Datos simplificado.

Con el modelo simplificado del Cubo de Datos se determina el modelo final en donde serán anotados semánticamente los datos obtenidos del REDI. En el modelo se identifica el conjunto de datos, el valor a ser observado que representa además la medida del Cubo de Datos y las dimensiones correspondientes.

Capítulo 4

Transformación de datos semánticos a datos dimensionales

En este capítulo se describe el proceso práctico realizado para organizar los datos sobre un modelo multidimensional, además se menciona las herrameintas utilizadas para realizar dicho proceso y finalmente se explica el Cubo de Datos resultante. Es decir en este capítulo se explica los procesos realizados para cumplir las etapas 2 y 3 del modelo de arquitectura de transformación de datos indicado en el capítulo anterior.

4.1. Introducción

Los datos obtenidos del sistema del REDI han pasado por un proceso de selección de datos y representación de instancias haciendo uso de vocabulario del Cubo de Datos, de esta manera se han descrito las relaciones que existen entre los dos formatos de datos semánticos y a partir del proceso de mapeo se obtuvo el modelo simplificado de la estructura del Cubo de datos. Continuando con la descripción del modelo de arquitectura (Ver Figura 3.1) que representa el proceso de transfomación de datos, en los siguientes apartados se describe las etapas siguientes.

4.2. Transformación de Datos en RDF a un Cubo de Datos

Conociendo los datos que estan disponibles dentro del REDI mediante el desarrollo de las etapas anteriores (Etapa 1) del proceso descrito dentro de la arquitectura propuesta(Ver Figura 3.1) se inicia con la segunda etapa que corresponde al análisis de los datos que serán organizados en una estructura multidimensional y la anotación semántica dentro de la escturctura mencionada, es decir, los datos extraídos deben ser integrados a la ontología del Cubo de Datos.

4.2.1. Definición de la estructura de la ontología del Cubo de Datos

Para definir la ontología en donde se instanciarán los datos en RDF, se utilizó el modelo multidimensional simplificado descrito en la Figura 3.5, en donde se define los valores correspondientes al Vocabulario del Cubo de Datos (Ver Figura 2.7) estos valores fueron obtenidos a través del mapeo de la ontología del REDI y la ontología del Cubo de Datos, la ontología está compuesta por clases y propiedades, en donde las clases definen la organización que corresponde a las dimensiones y medidas, y en las propiedas de la ontología se instancian los datos que son extraídos desde la plataforma del REDI.

Como se mencionó en el capítulo anterior, los datos extraídos son: publicaciones, autor, coautor, palabra clave y fuente, dentro de la estructura de la ontología se definen las propiedades correspondientes a cada dimension definida en la clase **qb:DimensionProperty** (Ver Tabla 3.2), para asignar cada dimensión a la instancia mencionada del Vocabulario del Cubo de Datos, estas dimensiones deben ser llamadas por variables que identifiquen el valor que se va ha almacenar. Las dimensiones y la medida, conjuntamente con las variables de cada propiedad dentro de la ontología, están descritas de la siguiente manera:

- **Dimensiones:**
  - Autor $\Longrightarrow refAutor$
  - Coautor $\Longrightarrow refContribuidor$
  - Palabra Clave $\Longrightarrow refKeyword$
  - Fuente $\Longrightarrow refFuente$

- Periodo $\Longrightarrow refPeriodo$

- **Medida:**
  - Publicación $\Longrightarrow cpublicacionRedi$

Con las instancias que se identificó a través del mapeo de datos se procede a utilizar dichos elementos para formar la ontología en donde se instanciarán los datos, para realizar estas acciones dentro de la ontología se utilizó la herramienta *Protege*, editor manual de ontologías y framework para base de conocimientos, la cual facilitó la adecuación de la ontología y la instanciación de elementos, además permite obtener la representación gráfica de la ontología en donde serán anotados los datos multidimensionales, la ontología se muestra en la Figura 4.1.

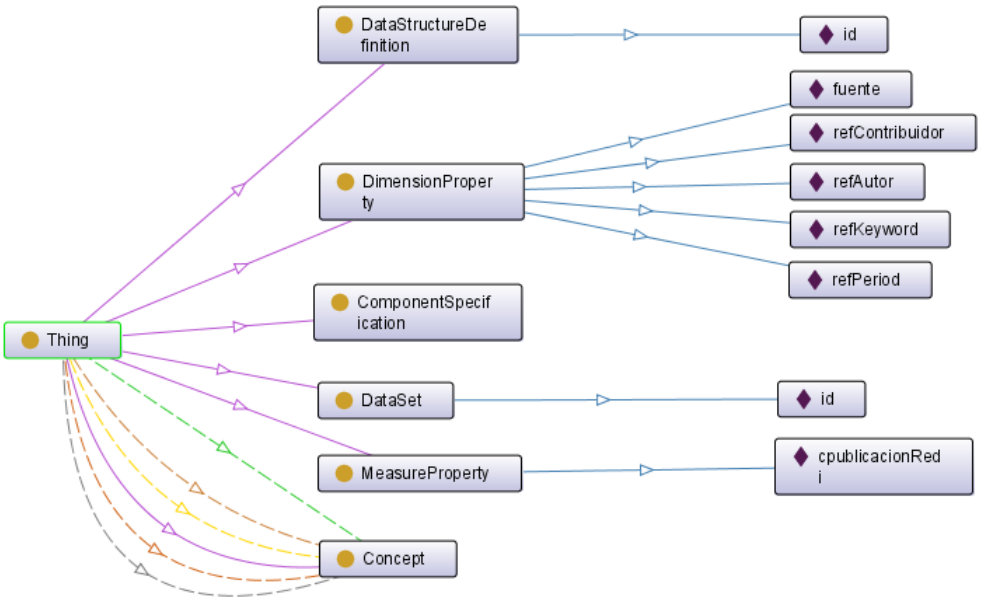

Figura 4.1: Estructura de la Ontología del Cubo de Datos

Como se puede apreciar en la ontología de la Figura 4.1, esta consta de las instancias o clases que se identificaron en el modelo simplificado del Vocabulario del Cubo de Datos que se presentó en el capítulo 3.

4.2.2. Instanciación de datos RDF en la Ontología del Cubo de Datos

La instanciación se lleva a cabo al anotar semánticamente los datos dentro de una estructura multidimensional, es decir, que los datos que estan escritos en el grafo central con formato RDF deben ser ahora escritos sobre un almacén de datos con estructura multidimensional, este proceso se automatizó desarrollando una aplicación que permita la interacción entre los datos del sistema del REDI y el almacén de datos multidimensional.

En la Figura 4.2 se indica el diagrama de flujo de la aplicación desarrollada para realizar el cambio en la estructura de los datos. Para desarrollar la aplicación se utilizó la herramienta NetBeans conjuntamente con la librería Jena, ya que esta librería permite la interacción directa entre el modelo ontológico y el almacén de datos del REDI.

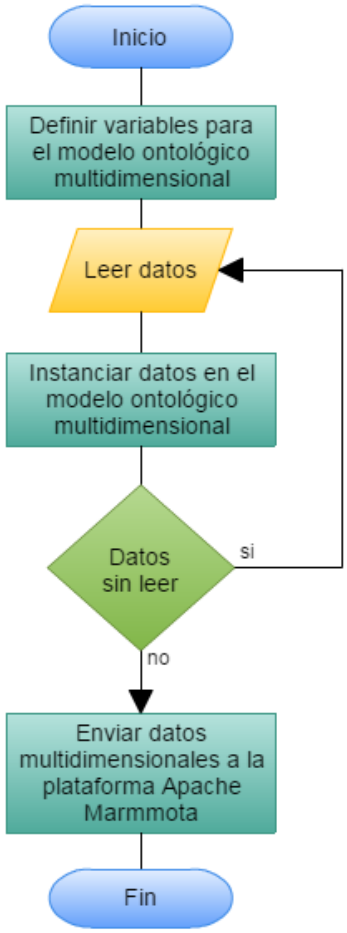

Figura 4.2: Diagrama de flujo del algoritmo de transformación a datos multidimencionales.

Cada uno de los pasos descritos en el proceso de automatizar la transformación de datos, se explican con mayor detalle en los siguientes apartados:

1. **Definir variables para el modelo ontológico multidimensional:** En este paso se identifica las variables necesarias para relacionar las URIs de los recursos que forman parte de la ontología del Cubo de Datos. Con estas variables se hará mención a los recursos de la ontología, y de esta manera se define la interación de la aplicación con la ontología del Cubo de Datos. Mediante la librería Jena se puede leer y crear archivos RDF desde java, por lo tanto haciendo uso de esta librería estas variables obtendran las tripletas en RDF dentro de java, además de los valores finales que deben ser integrados en la estructura multidimensional.

2. **Leer Datos:** Este paso hace referencia al acceso que la aplicación tiene a la plataforma Apache Marmotta la cual contiene los datos del sistema REDI, la extracción se realiza mediante una consulta SPARQL (Ver Figura 3.3), con la cual se obtiene las publicaciones con sus respectivos autores, contribuidores, palabras clave, fuente y año de publicación. Para leer los datos dentro de la plataforma Apache Marmmota, en el desarrollo de la aplicación se debe incluir la conexión a dicha plataforma.

3. **Instanciar datos en el modelo ontológico multidimensional:** En esta paso se asignan los valores extraídos de la consulta realizada en el paso 2 a las variables establecidas en el paso 1. En este paso se desarrolla la anotación o escritura semántica de las tripletas obtenidas de la ontología en RDF, cada valor que esta representado en los nodos del grafo central se almacena dentro de la dimensión que a sido asignada en la distribución de variables para identificar los recursos y almacenarlos dentro de la estrcutura multidimensional. Por ejemplo, para almacenar la publicación, esta se almacena conjuntamente con un valor numérico que serviría para representar los datos de manera estadística haciendo referencia a las publicaciones ya que este elemento se a asignado como medida o como valor observado según define el vocabulario del Cubo de Datos. Este paso es repetitivo dependiendo el núemro de tripletas que han sido recuperadas al ejecutar la consulta de la Figura 3.3.

4. **Enviar datos multidimensionales a la plataforma Apache Marmotta:** En este paso se establece la conexión entre la aplicación y la plataforma Apache Marmmota para enviar los datos multidimensionales hacia un grafo creado en la plataforma, este grafo se denomina DataCube.

4.3. Cubo de Datos resultante del proceso de transformación

Es punto describe la tercera etapa dentro del modelo de arquitectura de transformación de datos, por lo que se describe la organización de los datos luego de haber cumplido su etapa de transformación. El resultado de la transformación de datos en RDF a datos multidimensionales han sido migrados a la plataforma Apache Marmmota utulizada por el sistema del REDI, estos datos que estan organizados bajo un formato multidimensional deben ser visualizados para el interés del usuario final. La figura 4.3 muestra como estan definidas las dimensiones en el Cubo de Datos, esta forma gráfica de visualizar la organización de los datos ayuda a distinguir con mayor claridad la estructura de los datos, además se aprecia que un Cubo de Datos puede contener un conjunto de cubosmas pequeños. En caso la medida "publicación"se va formando cubos más grandes que representan publicaciones asociadas al cubo principal a tres dimensiones: autor, tiempo y fuente, y a su vez se los asocia con otros cubos que representan otras dos dimensiones que son palabras claves y coautores.

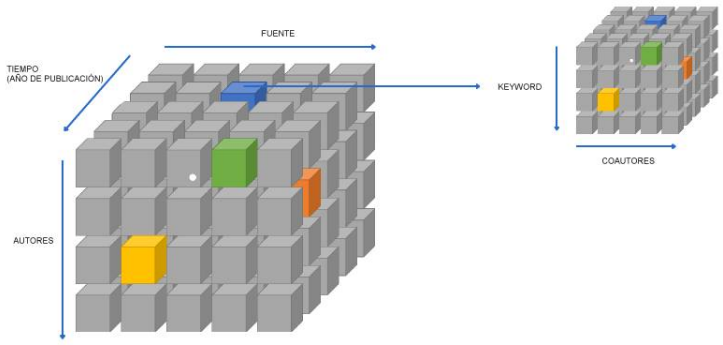

Figura 4.3: Representación gráfica del Cubo de Datos acoplada al proyecto de publicaciones

Una vez instanciados los datos, en cada uno de los cuadrados se encuentra una publicación

diferente. Por ejemplo, como se ve en la Figura 4.4 se puede apreciar dos ejemplos de publicaciones:

- **Publicación:** Late Pleistocene and holocene activity of the volcanic complex, publicado en el año 2008 con las palabras clave Ecuador, Holocene, Stratigraphy, Volcanic Hazards y con sus coautores, los mismos que se encuentran registrados con códigos en el servidor Marmmota.

- **Publicación:** Consuming and producing linked open the case of opencourseware del Autor Nelson Piedra, publicado en el año 2014 con las palabras clave: OpenCourseWare, Latin America, Floods y Forest.

Para tener una visión mas clara de como estan organizados los datos por cada dimensión establecida a través del mapeo de datos, se muestra en la Figura 4.4 ya que además de presentar las publicaciónes se presenta un ejemplo de datos y como podemos acceder a ellos.

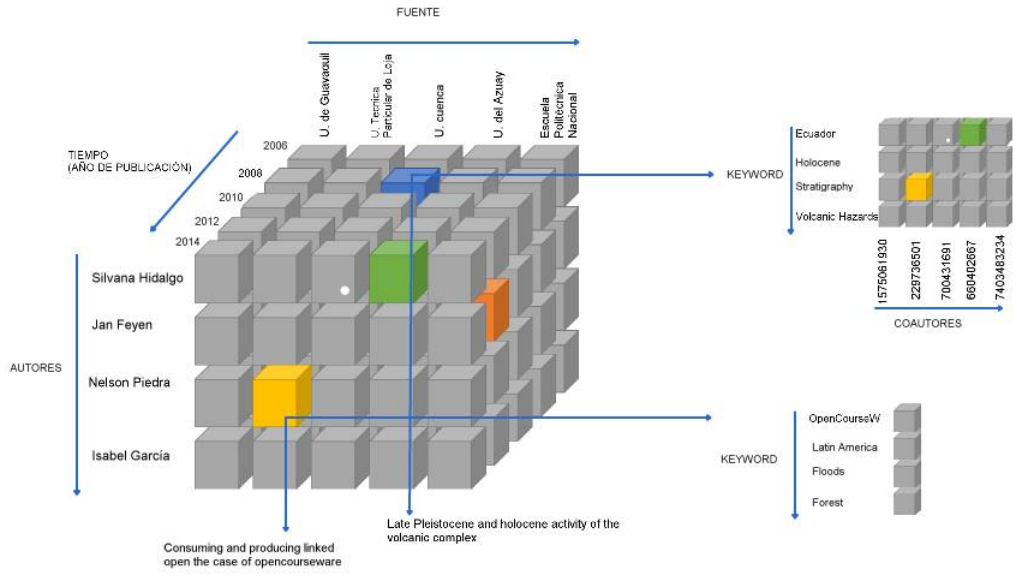

Figura 4.4: Organización de las publicaciones en el Cubo de Datos.

Capítulo 5

Visualización del Cubo de Datos

Debido a la necesidad de presentar al usuario la información de manera resumida y estadística se han desarrollado varias herramientas de visualización de información; dentro del proyecto se optó por la herramienta OpenCube Toolkit [27], la misma que mediante la utilización de Widgets preestablecidos facilita al desarrollador la adecuación de la interfaz para una mejor interpretación de la información por el usuario. En este capítulo se muestra de una manera gráfica los resultados obtenidos a partir del cubo de datos resultante.

5.1. Definición de la interfaz principal del visualizador

Previo a la visualización se realizó la importación de los datos, como se especifica en el Apéndice B, en el que mediante una consulta SPARQL se especifica los datos a ser importados junto con el grafo en el que se encuentran las tripletas, también permite definir el intervalo de tiempo en minutos para su actualización y cuenta con otras funcionalidades como acciones a realizar antes o después de la importación o definir restricciones para su importación, aunque para este trabajo de titulación solo se han utilizado las mencionadas. Posterior a este paso se creó una wiki denominada *PublicacionesDC* y a su vez fue acoplada de tal manera que se pueda demostrar el cumplimiento de los objetivos del proyecto de titulación planteados inicialmente.

Como se muestra en la Figura 5.1, esta Wiki está compuesta por un menú de cuatro opciones:

- **Estructura del Cubo de Datos:** Esta opción del menú está compuesta por dos secciones: la primera tiene una descripción general del proyecto y la segunda permite la visualización de la estructura del Cubo de Datos resultante del proceso de transformación.

- **Resumen Estadístico de Publicaciones:** En esta sección se visualiza la cantidad de publicaciones por autor, palabra clave y año de publicación, siendo estas las dimensiones especificadas en el Cubo de Datos. Estos datos son presentados en tablas y en Pie Charts, además cada una de ellas cuenta con un botón que redirecciona hacia una Wiki que permite visualizar los 10 autores, palabras clave y años de publicación que tengan el mayor número de publicaciones.

- **Búsqueda Dinámica de Publicaciones:** En esta Wiki se presenta una sección que contiene las dimensiones del Cubo de Datos, en donde permite seleccionar las que el usuario desea visualizar. La información se presentará en una tabla que muestra la cantidad de publicaciones por cada dimensión.

- **Información General:** Contiene Información general sobre el Cubo de Datos como: Estadísticas, clases, propiedades y jerarquías de la ontología; propiedades por rango y dominio; y relaciones de los datos en tripletas.

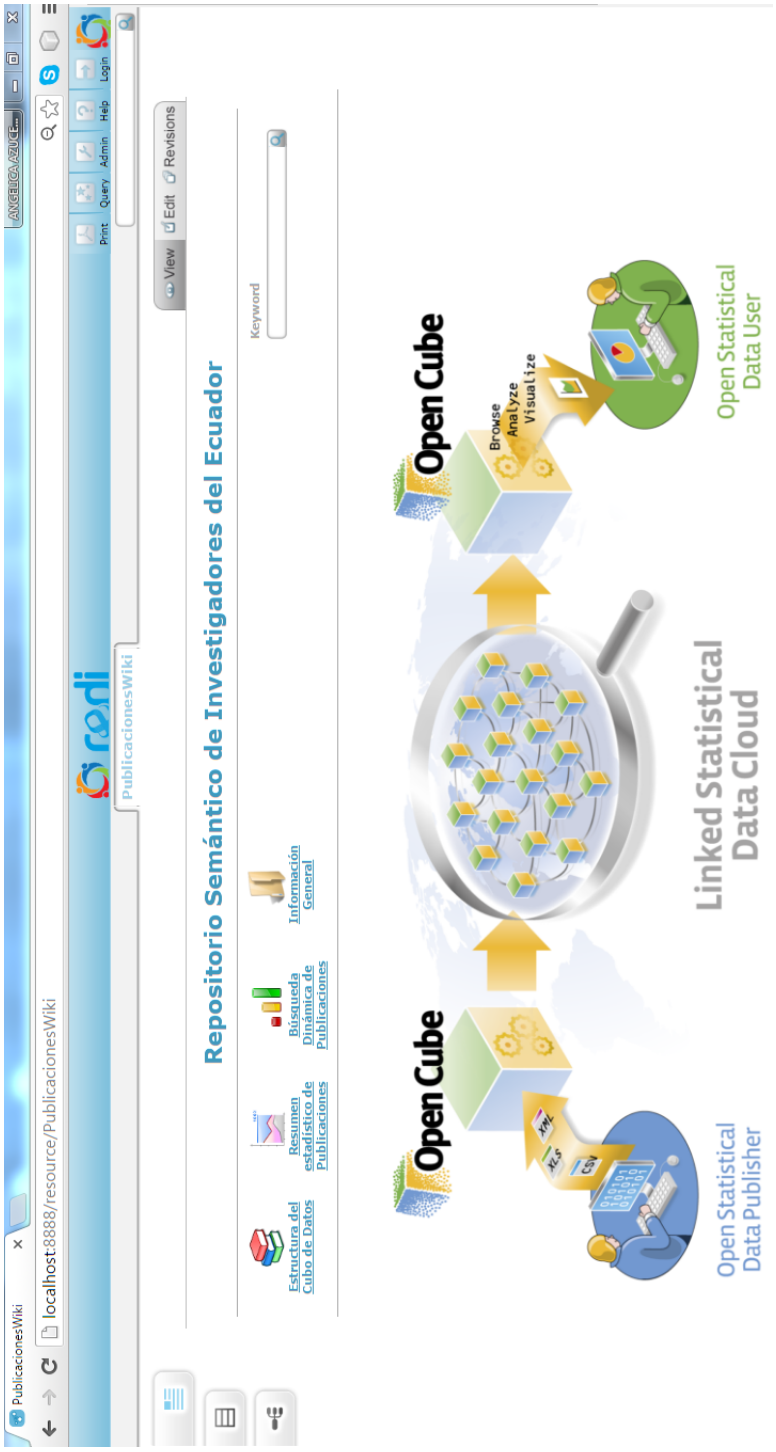

Figura 5.1: Wiki Principal.

5.1.1. Estructura del Cubo de Datos

Esta Wiki consta de dos secciones como se mencionó anteriormente, la primera con una descripción de la vinculación del actual proyecto con el sistema REDI en el que se define los objetivos principales del proyecto, las formas en las que se va a presentar la información en OpenCube Toolkit y las dimensiones consideradas; y la segunda con la estructura general del Cubo de Datos, presentada en forma de jerarquía. Esta wiki se presenta en la Figura 5.2. Además en la parte superior de cada wiki se presenta el menú para acceder a cualquiera de las cuatro opciones descritas al inicio del capítulo.

Figura 5.2: Wiki Estructura del Cubo de Datos.

En la estructura presentada en la Figura 5.2 se puede ver la jerarquía de los componentes del Cubo de Datos que a su vez está organizada por niveles cuyos primeros niveles son medidas, atributos, dimensiones y componentes. La medida del objeto de estudio en este proyecto es la publicación, la misma que se encuentra en el cubo de datos con su título y como propiedad numérica con el valor de uno para futuras contabilizaciones de publicaciones por cada dimensión; las dimensiones consideradas en este proyecto de titulación son: fuente de la publicación

como por ejemplo la Universidad de Cuenca, Universidad Técnica Particular de Loja, Universidad Politécnica Saleciana, entre otros; palabras clave que suelen ser palabras del título; año de publicación en el caso de tener, ya que algunas publicaciones no cuentan con esa información actualmente; autor y contribuidor. Los componentes que se visualiza en la jerarquía son las observaciones; y el atributo es la unidad de medida, que en este proyecto es numérico para el registro de la cantidad de publicaciones.

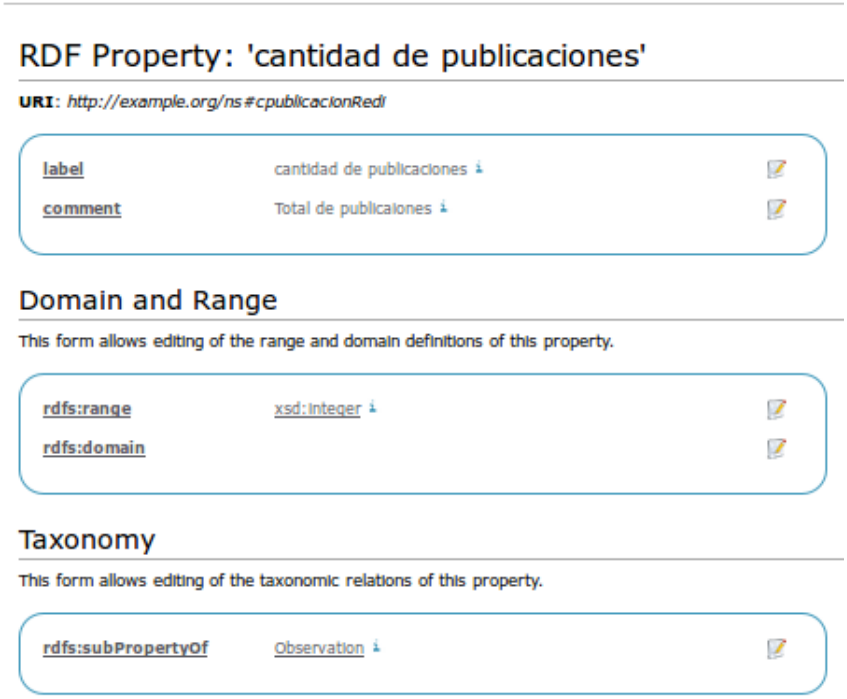

Figura 5.3: Detalle de la medida *Cantidad de Publicaciones*.

Cada uno de los enlaces redirecciona a una Wiki que detalla lo siguiente: propiedades, rango, dominio, taxonomía, subpropiedades y tripletas asociadas al recurso seleccionado. En el caso de querer visualizar la medida *cantidad de publicaciones*, se presenta una interfaz como se encuentra en la Figura 5.3 que a su vez presenta lo siguiente:

- Label: cantidad de publicaciones,
- Comment: total de publicaciones,

- Rango: entero y en taxonomía,
- Subpropiedad: observación y
- Tabla con títulos de publicaciones y su medida que es la cantidad de publicaciones con ese título, los mismos que se establecieron en el proceso de transformación, que en todos los casos con está medida, es uno.

Triples with predicate cantidad de publicaciones

1 - 30 / 200 Show 30 rows (max. 1000) Filter

subject	object
menopausal-symptoms-appear-before-the-menopause-and-persist-years-detailed-analysis-of-multinational-study	1
anthropometric-model-for-the-prediction-of-appendicular-skeletal-muscle-mass-in-chilean-older	1
spatial-and-temporal-trends-of-pcdds-and-pcdfs-in-bivalve-mollusc-coming-from-galicia-possible-relationship-between-biometric-parameters-and-pcdds-and-pcdfs-levels	1
consuming-and-producing-linked-open-the-case-of-opencourseware	1
multinational-study-of-sleep-disorders-during-female-mid-life	1
an-index-for-delivery-of-ecosystem-service	1
laparoscopic-resectional-gastric-bypass-in-patients-with-morbid-experience-on-112-consecutive-patients	1
subduction-zones	1
anthocyanins-and-related-detecting-the-change-of-regime-between-rate-control-by-hydration-or-by-tautomerization	1

Figura 5.4: Tripletas de la medida *Cantidad de publicaciones*.

Para el enlace de referencia fuente, se presenta de la misma manera, los recursos relacionados con la propiedad *referencia fuente*, que para este caso son: label: referencia fuente, comment: Fuente de la publicación, rango: Concept y subpropiedad: fuente, como se muestra en la Figura 5.5; y en una tabla se presenta las tripletas relacionadas con la propiedad *referencia fuente* que para este caso son los títulos de las publicaciones con el nombre de la fuente como se muestra en la Figura 5.6.

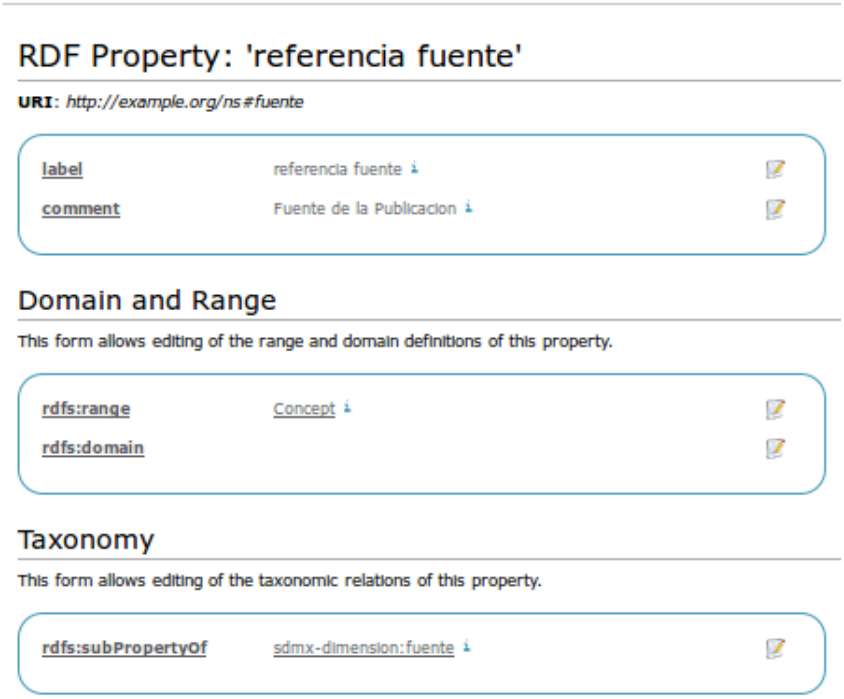

Figura 5.5: Detalle de la dimensión *Fuente*.

De la misma manera ocurre para el enlace referencia keyword, referencia periodo, referencia autor y referencia contribuidor, en donde se presentarán los recursos asociados a estas dimensiones en la ontología; y en las tripletas se visualiza los sujetos y objetos que tengan como predicado la dimensión seleccionada.

5.1.2. Resumen estadístico de publicaciones

En esta wiki se presenta un resumen de la cantidad de publicaciones por autor, fuente, palabras clave y año en tablas que listan cada una de las dimensiones con la cantidad de publicaciones y en Pie Charts que facilitan la visaulización general de las dimensiones con la cantidad de publicaciones como se muestra en la Figure 5.7.

Como se puede ver, cada uno de los resúmenes de publicaciones por dimensiones tiene un botón que redirecciona hacia los 10 atributos de cada dimensión con mayor cantidad de publicaciones. Además permite su visualización estadística por tres tipos de gráficos: BarChart, LineChart y PieChart.

Triples with predicate referencia fuente

1 - 30 / 200 Show 30 rows (max. 1000) Filter

subject	object
subduction-zones	EscuelaPolitecnicaNacional
petrological-analysis-of-the-pre-eruptive-magmatic-process-prior-to-the-2006-explosive-eruptions-at-tungurahua-volcano	EscuelaPolitecnicaNacional
can-retrieval-of-information-from-citation-indexes-be-multiple-mention-of-reference-as-characteristic-of-the-link-between-cited-and-citing-article	EscuelaPolitecnicaNacional
sentiment-analysis-of-citations-using-sentence-structure-based-features	EscuelaPolitecnicaNacional
revision-of-the-south-american-nematognathi-or-cat-fishes	EscuelaPolitecnicaNacional
the-colima-volcanic-post-caldera-andesites-from-volcan-colima	EscuelaPolitecnicaNacional
adakitic-magmas-in-the-ecuadorian-volcanic-petrogenesis-of-the-iliniza-volcanic-complex	EscuelaPolitecnicaNacional
nitrogen-and-argon-isotopes-in-oceanic-basalts	EscuelaPolitecnicaNacional
late-pleistocene-and-holocene-activity-of-the-volcanic-complex	EscuelaPolitecnicaNacional

Figura 5.6: Tripletas de la dimensión *Fuente*.

Por ejemplo, al dar click en el primer boton: *Visualizar 10 autores con mayor cantidad de publicaciones*, se muestra una wiki con un menú en la parte superior izquierda, que permite visualizar los datos en una tabla como se muestra en la Figura 5.8 y por tres tipos de gráficos diferentes: BarChart, LineChart y PieChart como se muestra en las figuras: 5.9, 5.10 y 5.11.

AUTORES Y CANTIDAD DE PUBLICACIONES

1 - 10 / 33 Show 10 rows (max. 1000)
Filter

type	sumX
isabel-susana-garcia-valdez	16
hugo-sánchez-romero	11
maria-espinoza	17
carlos-castro-riera	12
jorge-luis-rojas-vivanco	12
silvana-ivonne-hidalgo-trujillo	24
maria-elisa-ordoñez-alvarado	6
peter-l-m-goethals	15
jan-feyen	15
felipe-cisneros	10

Visualizar 10 autores con mayor cantidad de publicaciones

GRÁFICA

PUBLICACIONES POR FUENTES

1 - 10 / 22 Show 10 rows (max. 1000)
Filter

type	totalpublicaciones
UniversidaddelAzuay	2
http://190.15.141.66:8899/uce/	2
UniversidadEstataldeBolivar	2
UniversidadPolitecnicaEstataldeCarchi	2
UniversidadSanFranciscodeQuito	2
UniversidaddeGuayaquil	2
http://190.15.141.66:8899/puce/	2
UniversidadTecnicadelNorte	2
UniversidadCatolicadeSantiagodeGuayaquil	2
http://190.15.141.66:8899/espoch/	2

Visualizar 10 fuentes con mayor cantidad de publicaciones

GRÁFICA

PUBLICACIONES POR PALABRAS CLAVE

1 - 10 / 544 Show 10 rows (max. 1000)
Filter

type	totalpublicaciones
ES	2
PCBs	4
Factorial design	2
Mytilus galloprovincialis	2
Accelerated solvent extraction	2
GC-MS-MS	2
Climacteric Symptoms	2
Blood Pressure	2
Menopause	9
Latin America	15

Visualizar 10 palabras claves con mayor cantidad de publicaciones

GRÁFICA

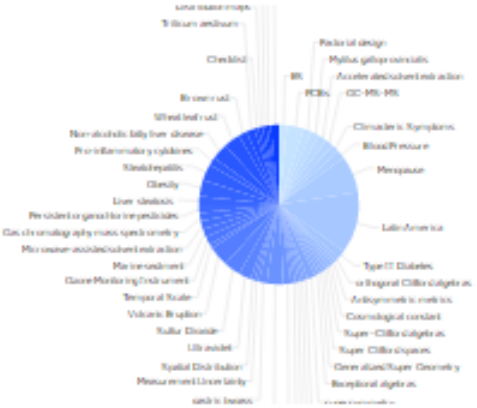

Figura 5.7: Wiki: Resumen estadístico de Publicaciones.

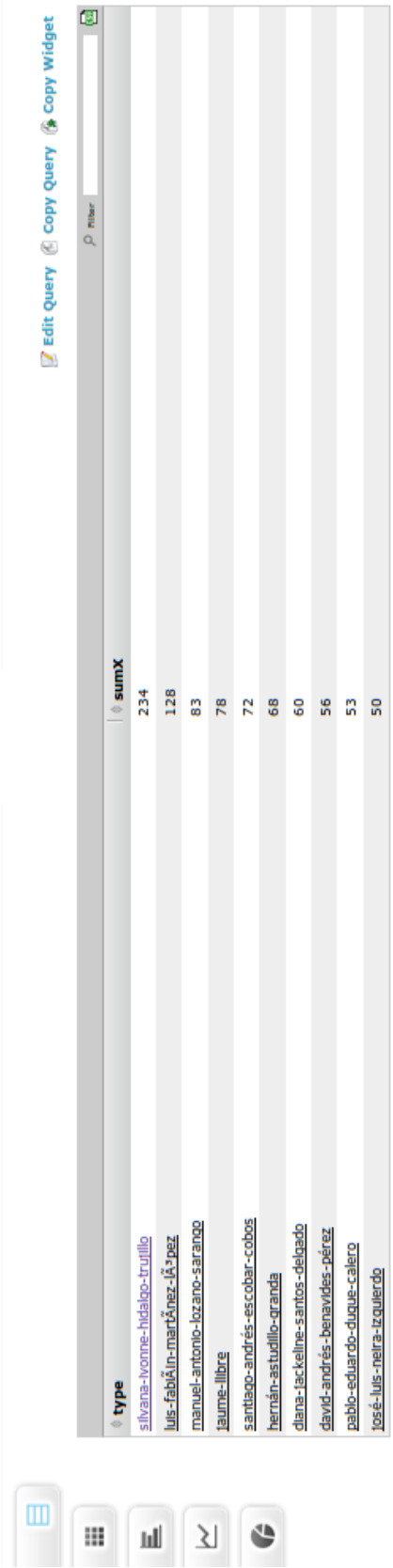

Figura 5.8: 10 autores con la mayor cantidad de publicaciones.

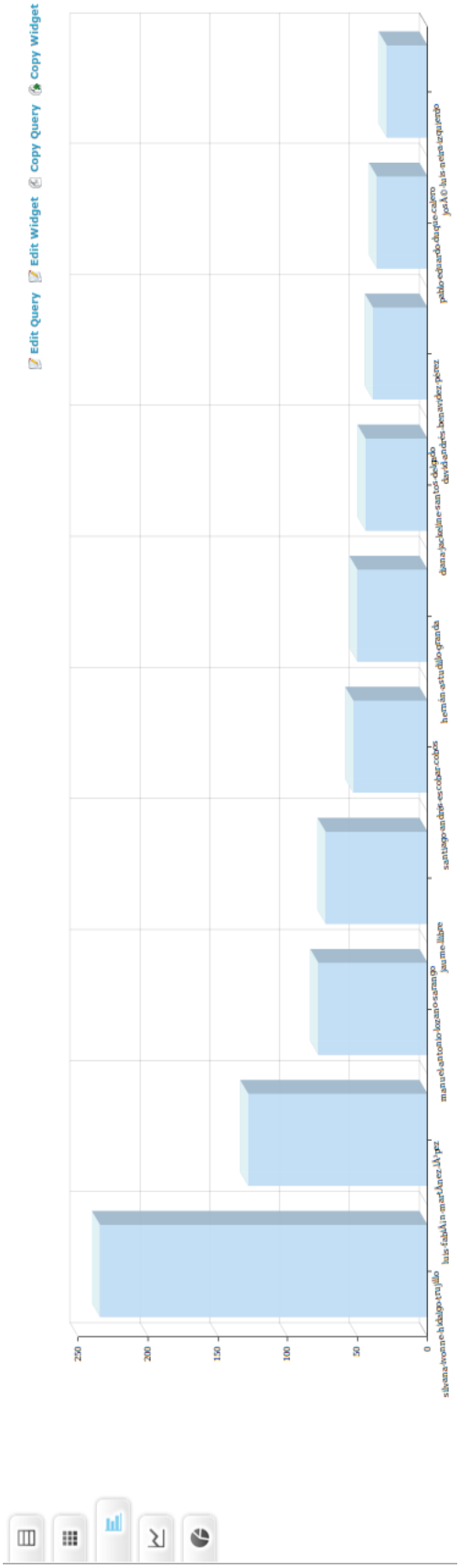

Figura 5.9: Visualización gráfica – BarChart.

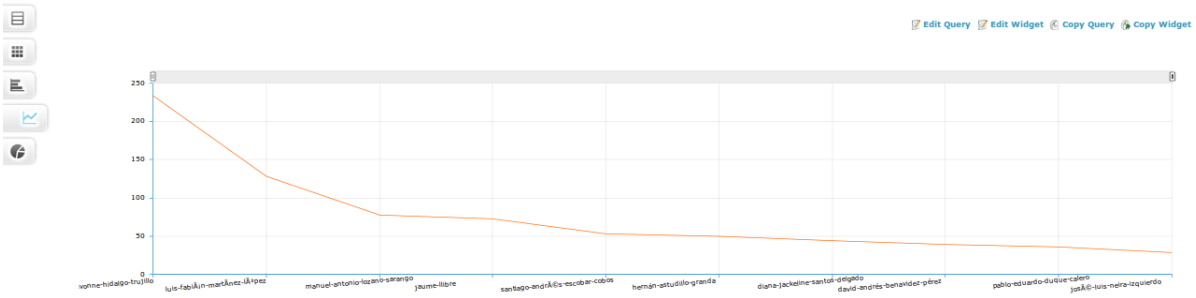

Figura 5.10: Visualización gráfica – LineChart.

Además al seleccionar alguno de los datos presentados en las tablas presentadas en la Figura 5.7 o 5.8, se muestra otro menú que facilita la exploración del cubo de datos como se muestra en la Figura 5.12 que contiene tres opciones:

- Wiki view: En esta wiki se puede personalizar por Uri existente en el cubo de datos. Por ejemplo, al seleccionar la autora Susana García, no se visualiza nada ya que no existe una wiki preestablecida con información relacionada, sin embargo, en un futuro se puede personalizar y enriquecer la ontología con mayor cantidad de recursos para visualizar en esta wiki.

- Table view: En esta sección se muestra una wiki que contiene los recursos y sus vínculos de acceso relacionados al recurso seleccionado. Por ejemplo, al seleccionar la autora Susana García, se visualiza sus publicaciones y las publicaciones con sus coautores mediante los cuales se puede acceder a la autora como se muestra en la Figura 5.13.

- Graph view: Wiki que contiene los Recursos relacionados a la Uri seleccionada. Por ejemplo, el grafo que se presenta en la Figura 5.14 es la representación de las Uris relacionadas con la Uri de la autora Susana García en el cubo de datos. Además se puede acceder a cada uno de ellos y visualizar las Uris relacionadas con la misma como se muestra en la Figura 5.15 en donde se puede observar que la autora es Susana García, su fecha de publicación es en el 2008, la fuente de la publicación es la Universidad de Cuenca, los contribuidores, las palabras clave, el nombre del data set, entre otros atributos de la ontología que se encuentran relacionados con la Publicación *Spatial and temporal trends of pcdds and pcdfs*

in bivalve mollusc coming from galicia possible relationship between biometric parameters and pcdds and pcdfs levels de la autora Susana García.

De la misma manera se mostrará este menú que facilita y permite la exploración del cubo de datos con cualquier autor o contribuidor que se seleccione de las tablas presentadas en la Figura 5.7 o 5.8.

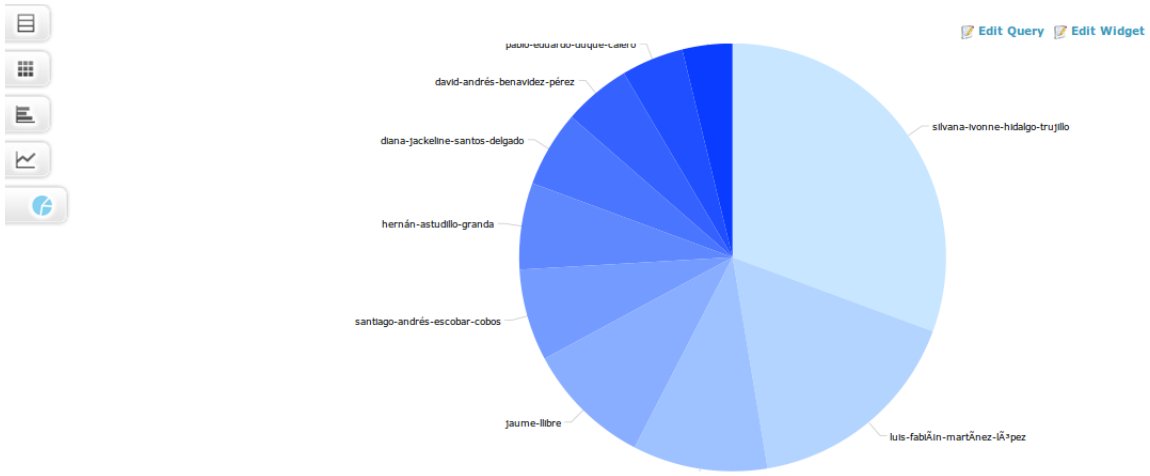

Figura 5.11: Visualización gráfica – PieChart.

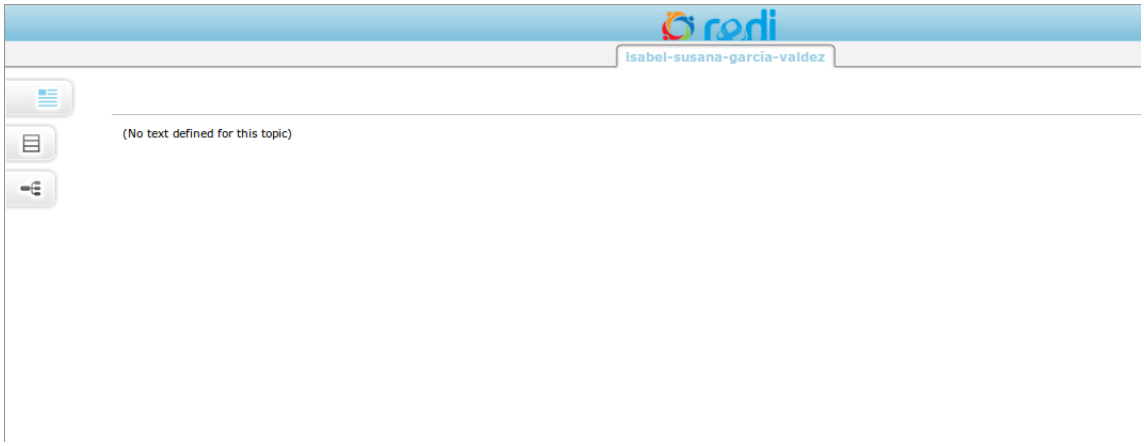

Figura 5.12: Vista en tabla de la autora Susana García.

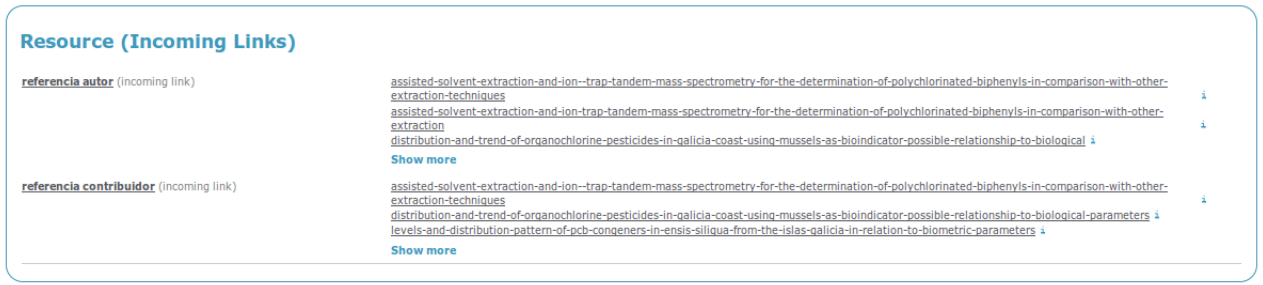

Figura 5.13: Vista en tabla de la autora Susana García.

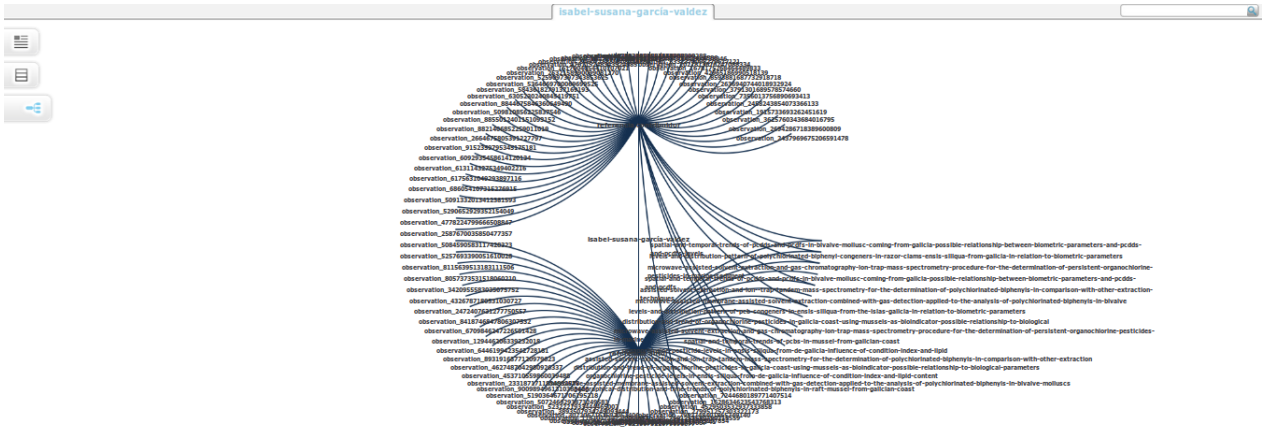

Figura 5.14: Opciones de menú para la autora Susana García.

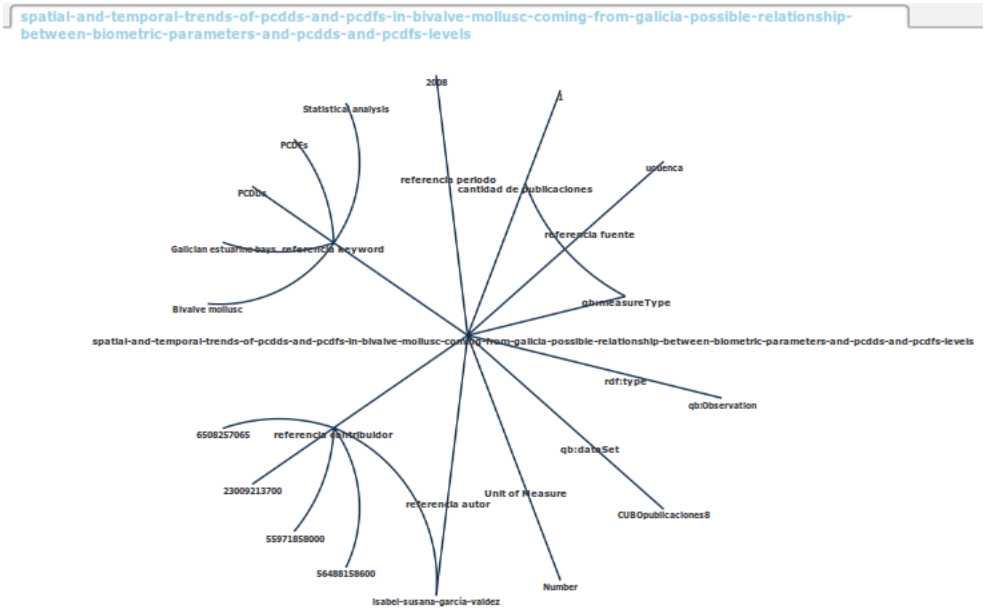

Figura 5.15: Vista por nodos de la publicación de la autora Susana García.

5.1.3. Búsqueda dinámica de publicaciones

En esta wiki se utilizó el widget browser mencionado en el capítulo 2 para presentar las dimensiones y medidas del cubo de datos. A su vez permite al usuario elegir las dimensiones que quiera visualizar; y al seleccionarlas podrá visualizar la cantidad de publicaciones correspondientes a las mismas.

Al seleccionar una dimensión se presenta la cantidad de publicaciones por cada instancia de la dimensión seleccionada en donde la primera columna presenta las instancias de la dimensión y en la segunda la medida. Por ejemplo, como se muestra en la Figura 5.16, la dimensión seleccionada es autor de publicación y la medida Total de publicaciones, los mismos que son presentados en una tabla con un paginamiento de 10 autores con sus respectivas cantidades de publicaciones.

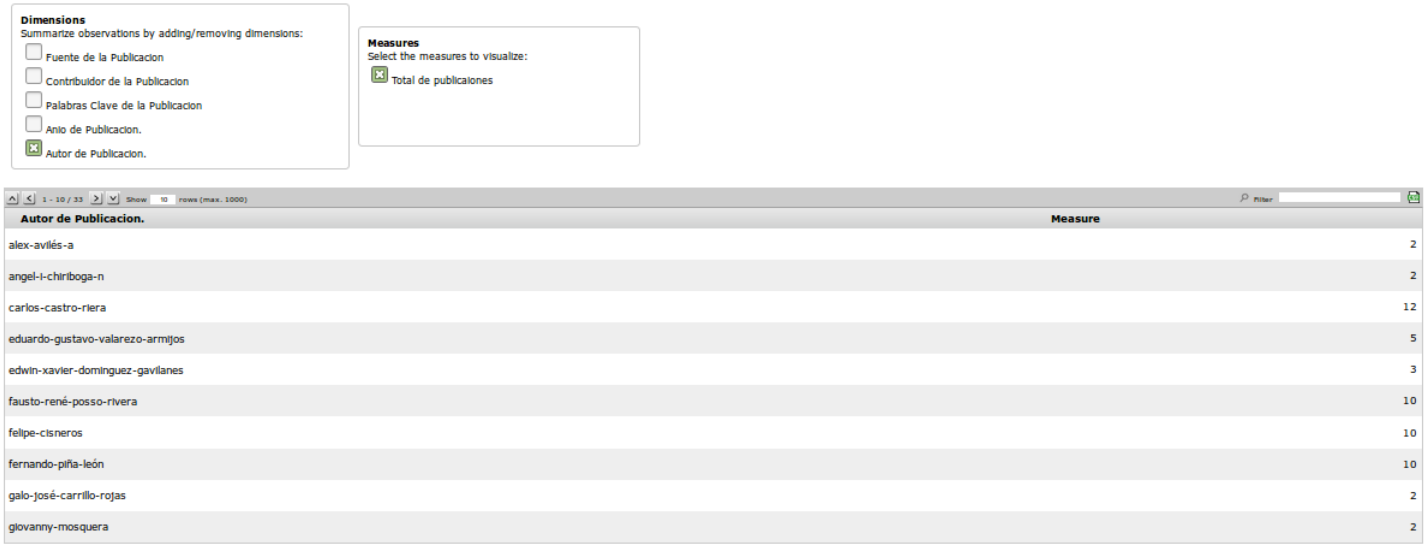

Figura 5.16: Cantidad de publicaciones por una dimensión.

Al seleccionar dos dimensiones se presenta la cantidad de publicaciones correspondientes a las dos dimensiones seleccionadas y un cuadro que permite escoger que dimensión visualizar como fila y que dimensión visualizar como cabecera de columnas. Por ejemplo, como se muestra en la Figura 5.17, las dimensiones seleccionadas son fuente de publicación y Autor de publicación con la medida Total de publicaciones y especificado en el cuadro inferior que la visualización de la tabla se va a realizar con la dimensión Fuente de publicación como cabecera de las columnas y la dimensión Autor de Publicación como filas o valores de la primera columna.

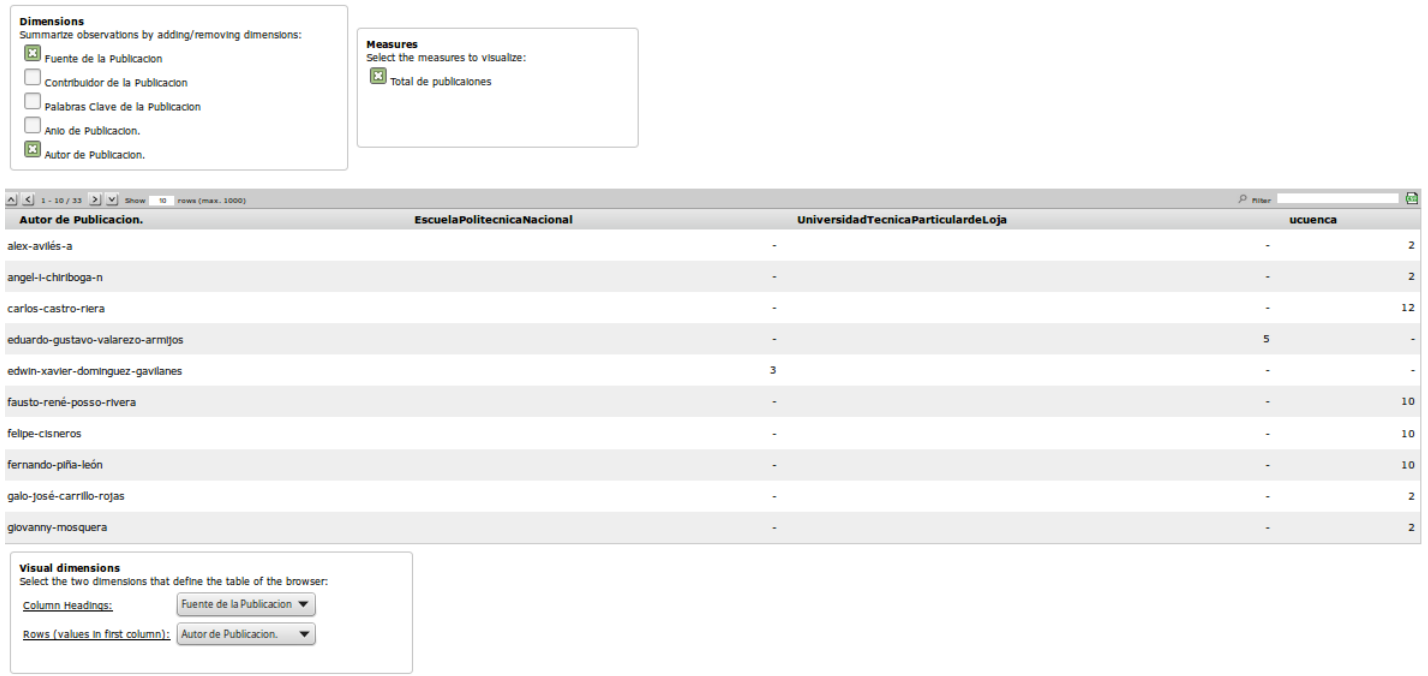

Figura 5.17: Cantidad de publicaciones por dos dimensiones.

Al seleccionar tres dimensiones se presenta la cantidad de publicaciones correspondientes a las dos dimensiones establecidas como fila y columna especificadas en el cuadro inferior con el total de publicaciones correspondientes y filtrado por un atributo específico de la tercera dimensión seleccionada. Por ejemplo como se muestra en la Figura 5.18 las dimensiones seleccionadas son: Fuente de Publicación, Año de publicación y Autor de Publicación, la medida es la cantidad de publicaciones, en donde la fila o valores de la primera columna y las cabeceras de las columnas a visualizar en la tabla son Autor y Fuente de publicación respectivamente y el filtro elegido para la dimensión Año de publicación es el año 2005; acorde a estos datos elegidos se puede observar que el autor Alex Aviles ha realizado dos publicaciones en el año 2005, Angel Chiriboga no ha publicado en ese año, Carlos Riera ha realizado 3 publicaciones en la Universidad de Cuenca en el año 2005, Edwin Dominguez ha realizado una publicación en la Universidad Técnica Particular de Loja, entre otros, además la visualización se realiza en un paginado de 10 autores.

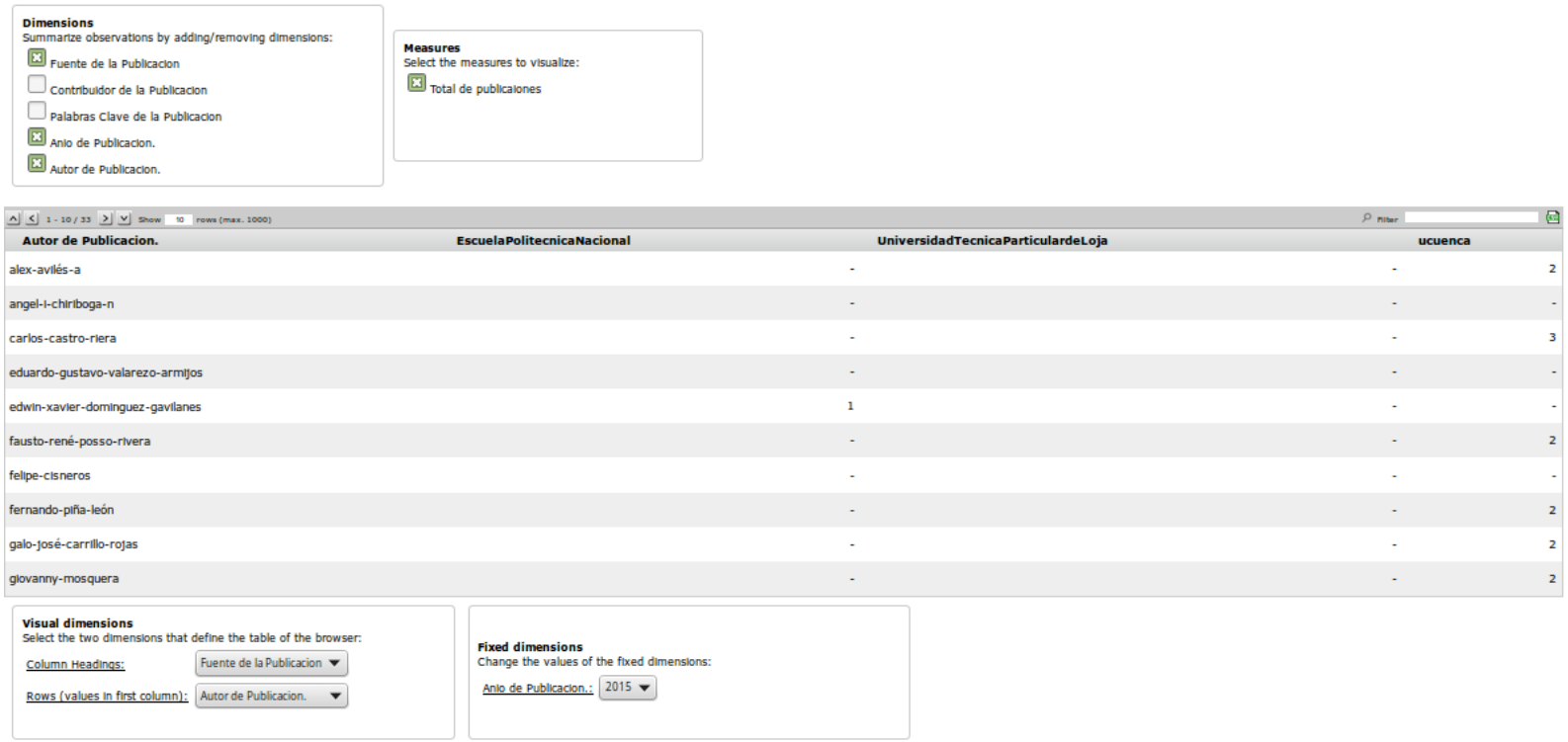

Figura 5.18: Cantidad de publicaciones por tres dimensiones.

5.1.4. Información General

En esta Wiki se presenta la información relacionada al Cubo de Datos importado inicialmente, la información se presenta en seis secciones, que son las siguientes:

- Estadística: Muestra la cantidad de tripletas, clases, propiedades, entidades; y sujetos y objetos distintos. Además, permite recalcular y eliminar las estadísticas. (véase la Figura 5.19).

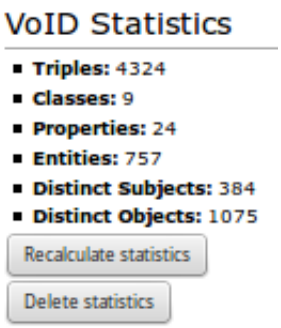

Figura 5.19: Información Estadística del Cubo de Datos.

- Clases: Muestra las clases definidas en el Cubo de Datos que se especificaron para este proyecto. Además, varían su tamaño dependiendo la cantidad de instancias en cada una. Como se observa en la figura 5.20, se visualiza que la clase observación es la más grande debido a que contiene la mayor cantidad de instancias.

Figura 5.20: Información de Clases del Cubo de Datos.

- Propiedades: Similar a la sección de las clases, también varían su tamaño y hacen referencia a las propiedades definidas en el Cubo de Datos inicialmente, en este caso se visualiza que las propiedades *referencia contribuidor* y referencia keyword son las más grades debido a que contienen una mayor cantidad de instancias (véase la Figura 5.21).

Figura 5.21: Información de Propiedades del Cubo de Datos.

- Jerarquía de Clases: En esta sección se muestra de manera jerárquica las clases definidas en el Cubo de Datos inicialmente (véase la Figure 5.22).

Figura 5.22: Información de la Jerarquía de Clases de Cubo de Datos.

- Propiedades por Rango y Dominio: Lista todas las propiedades por su Rango y Dominio (véase la Figura 5.23).

- Tripletas: En esta sección se muestran las tripletas que contiene el Cubo de Datos, con un límite de 10000 tripletas. (véase la Figura 5.24).

Properties by range and domain

The table below lists all properties by their range and domain.

Property	Domain	Range
rdf:type		
label		
rdfs:range		
qb:codeList		
qb:concept		
comment		
rdfs:subPropertyOf		
qb:dataSet		
Unit of Measure		
referencia autor		Concept
referencia contribuidor		Concept
referencia keyword		Concept
referencia periodo		Concept
cantidad de publicaciones		xsd:integer
referencia fuente		Concept
qb:measureType		
qb:structure		
aggregationSet		
qb:component		
qb:attribute		
qb:componentAttachment		
qb:componentRequired		
qb:measure		
qb:dimension		

Figura 5.23: Información de las Propiedades por Rango y Dominio del Cubo de Datos.

Triples

The following triples are contained in this context. Please note, that the number of triples retrieved is limited to 10,000.

1 - 30 / 4324 Show 30 rows (max. 1000)

s	p	o
qb:dataSet	rdf:type	AnnotationProperty
referencia autor	rdf:type	qb:DimensionProperty
referencia autor	rdf:type	rdf:Property
referencia autor	label	referencia autor
referencia autor	rdfs:range	Concept
referencia autor	qb:codeList	id
referencia autor	qb:concept	sdmx-concept:refAutor
referencia autor	comment	Autor de Publicacion.
referencia autor	rdfs:subPropertyOf	sdmx-dimension:refAutor
referencia contribuidor	rdf:type	qb:DimensionProperty
referencia contribuidor	rdf:type	rdf:Property
referencia contribuidor	label	referencia contribuidor
referencia contribuidor	rdfs:range	Concept
referencia contribuidor	qb:codeList	id
referencia contribuidor	qb:concept	sdmx-concept:refContribuidor
referencia contribuidor	comment	Contribuidor de la Publicacion
referencia contribuidor	rdfs:subPropertyOf	sdmx-dimension:refContribuidor
referencia keyword	rdf:type	qb:DimensionProperty
referencia keyword	rdf:type	rdf:Property
referencia keyword	label	referencia keyword
referencia keyword	rdfs:range	Concept
referencia keyword	qb:codeList	id
referencia keyword	qb:concept	sdmx-concept:refKeyword
referencia keyword	comment	Palabras Clave de la Publicacion
referencia keyword	rdfs:subPropertyOf	sdmx-dimension:refKeyword
referencia periodo	rdf:type	qb:DimensionProperty
referencia periodo	rdf:type	rdf:Property
referencia periodo	label	referencia periodo
referencia periodo	rdfs:range	Concept
referencia periodo	qb:codeList	id

Figura 5.24: Información de las Tripletas que contiene el Cubo de Datos.

Capítulo 6

Conclusiónes y Trabajos Futuros

6.1. Conclusión

En este capítulo se presentan los resultados que han sido cumplidos de acuerdo al planteamiento del objetivo principal para el desarrollo del presente trabajo de titulación. El objetivo fue el realizar un proceso para transformar los datos semánticos RDF a datos multidimensionales, además de visualizar los datos de manera estadística. A continuación se describen las principales actividades y objetivos que se cumplieron al culminar el desarrollo de este trabajo.

1. Para cumplir con los objetivos se utilizó el vocabulario del Cubo de Datos que propone la W3C, a este vocabulario se lo simplificó de manera que se adapte a la necesidades del objetivo planteado, con la adecuación se obtuvo una ontología que permitía que los datos tengan una estructura multidimensional.

2. A partir de la ontología se generó un Cubo de Datos mediante la aplicación desarrollada con la información del proyecto REDI que esta en la plataforma Apache Marmotta.

3. El cubo de datos con las instancias de las publicaciones fue importado a la herramienta OpenCube Toolkit para su visualización estadística.

4. En la visualización se presenta cuatro estructuras principales que son: la estructura de la ontología del Cubo de Datos, información estadística de cada una de las dimensiones con

la cantidad de publicaciones, visualización dinámica de las dimensiones con su cantidad de publicaciones y la información general del Cubo de datos con sus instancias.

5. Se ha incorporado el presente trabajo de titulación al proyecto REDI.

Finalmente, ya con el sitio online accesible por el público en general, se da por cumplido el objetivo de este trabajo de titulación, que es la transformación de los datos semánticos del proyecto REDI y su visualización estadística, pero además, dejando la puerta abierta a nuevos proyectos y mejoras.

6.2. Trabajos Futuros

Como propuesta de un trabajo que se podría realizar a futuro, es el desarrollo de una aplicación que permita pasar cualquier tipo de datos semánticos, en el que el usuario pueda definir el nombre de su cubo de datos, las dimensiones y las medidas que requiera, esto en la parte del back-end. En la parte del front-end, se puede establecer plantillas en la herramienta Open-Cube Toolkit que permita seleccionar los cubos de datos generados y así facilitar al usuario la visualización estadística del cubo de datos generado por la aplicación.

Apéndice A

Acronyms

RDF Resource Description Framework.

SPARQL Acrónimo recursivo del inglés SPARQL Protocol and RDF Query Language.

OLAP On-Line Analytical Processing.

W3C (World Wide Web Consortium) Es una comunidad internacional que desa- rrolla estándares que aseguran el crecimiento de la Web a largo plazo.

LDP (Linked Data Plataform) Describe un método de publicación de datos estructurados para que puedan ser interconectados y más útiles.

FOAF (Friend of a friend) Es una ontología legible para las máquinas que describe a las personas, sus actividades y sus relaciones con otras personas y objetos. Para hacer estas descripciones utiliza el Marco de Descripción RDF y el lenguaje de marcado OWL.

OWL (Ontology Web Language) Lenguaje de marcado para publicar y compartir datos usando ontologías en la WWW.

UML (Unified Modeling Language) Lenguaje de modelado de sistemas de software.

Apéndice B

Instalación de OpenCube Toolkit e importación del Cubo de Datos

A continuación se describe el proceso de instalación de la herramienta OpenCube Toolkit, así como las herramientas que se debe tener previo a la instalación y la importación del Cubo de Datos de las publicaciones.

B.1. Instalación de OpenCube Toolkit

Para correr OpenCube Toolkit se necesita:

- Una versión compatible de Java (preferiblemente la versión 1.7).
- La distribución de OpenCube Toolkit (archivo .zip).

Para comenzar la instalación se debe seguir los siguientes pasos:

- Descomprimir el archivo .zip de OpenCube Toolkit.
- Ubicarse en el directorio de OpenCube Toolkit descomprimido anteriormente.
- En el archivo iwb.sh dentro de la carpeta fiwb, modificar el valor para RUN_AS_USER=X en donde X representa el usuario.

- En el caso de linux, se debe ejecutar el sript linux-install.sh que se encuetra en el directorio donde se descomprimio la herramienta OpenCube Toolkit.
- Finalmente se ejecuta el scrip start.sh.

B.2. Importación del Cubo de Datos de Publicaciones en la herramienta OpenCube Toolkit

Para la importación del Cubo de Datos que se encuentra en la plataforma apache marmotta y poder visualizar estadisticamente la información, se realizó cuatro principales pasos:

- Crear un proveedor de datos desde un endpoint a través de una consulta Construct.
- Identificar si el cubo de datos es compatible.
- Calcular la suma de las publicacaciones mediante las dimensiones establecidas, este proceso.

B.2.1. Creación de proveedor

El primer paso fue crear un proveedor (véase la Figura B.1), el mismo que permite definir un nombre para el cubo de datos, el tiempo en el que deben actualizarse, usuario y password en caso de que la fuente de datos lo tenga, endpoint y la Query para extraer los datos. Adicional a lo establecido, se puede establecer reglas como límites o variables a importar pero para el presente trabajo de titulación no se necesitó.

B.2.2. Identificación de compatibilidad del cubo de datos

El segundo paso fue identificar si el cubo de datos que se obtuvo luego de extraer del repositorio del REDI y transformarlos a multidimensionales era válido, esta verificación se realizó mediante el widget OpenCube Compatibility Explorer (véase la Figura B.2). Una vez verificado

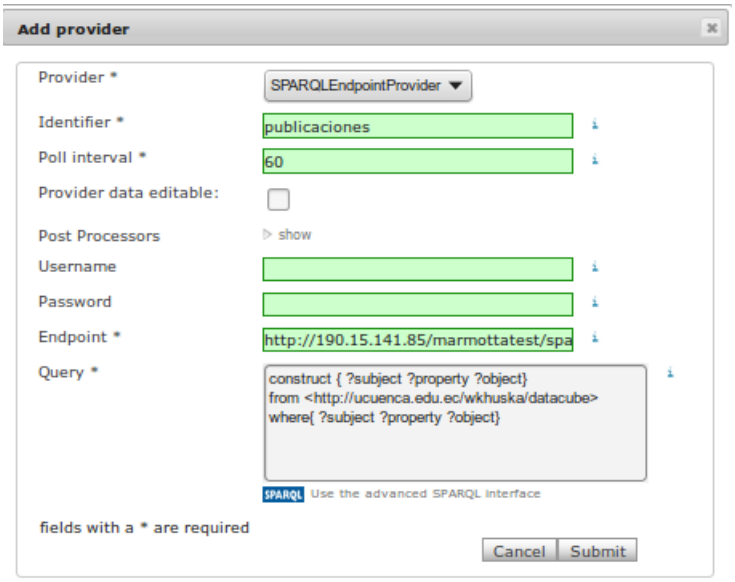

Figura B.1: Creación de proveedor de datos.

la validez del cubo de datos, se genera un link (véase la Figura B.3), el mismo que se utiliza en el siguiente paso.

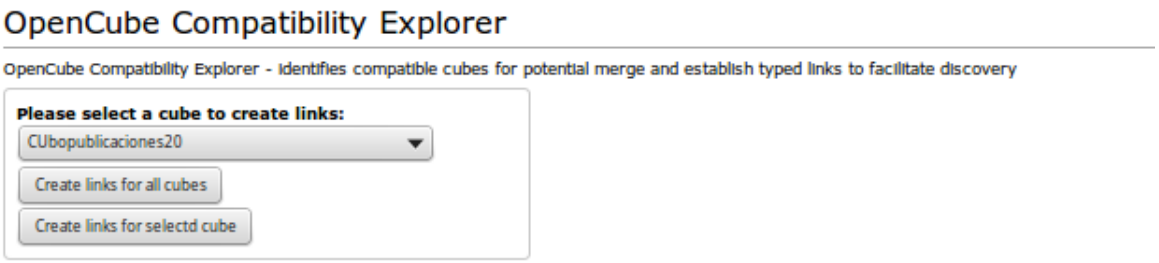

Figura B.2: Verificación del cubo de datos.

B.2.3. Cálculo de sumas por cada dimensión

El tercer paso fue calcular la suma de las publicaciones por autor, coautor, palabra clave y anio de publicación mediante el widget OpenCube Aggregator; en el que internamente, Open Cube Toolkit, crea $2^n - 1$ cubos en donde n, es la cantidad de dimensiones. (véase la Figura B.4).

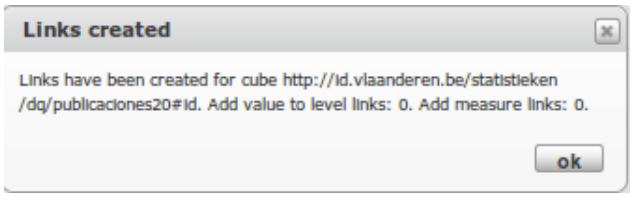

Figura B.3: Link verificado del cubo de datos.

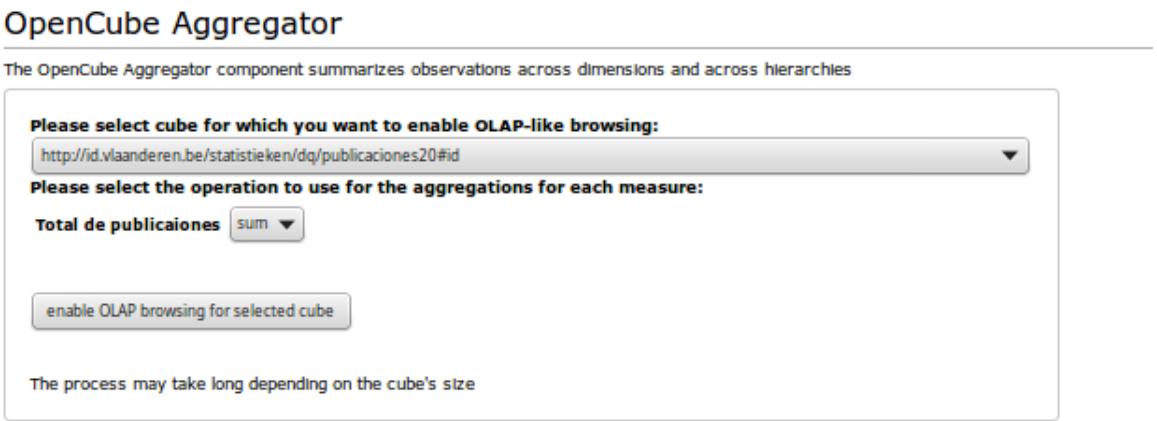

Figura B.4: Suma de publicaciones por dimensiones.

Bibliografía

[1] Apache Marmotta. *http://marmotta.apache.org/*.

[2] ARANO, Silvia. *http://eprints.rclis.org/9020/*. 2012.

[3] BRAY, Tim. *http://www.xml.com/pub/a/98/06/rdf.htm*. 2012.

[4] Francisco Javier Cruz Vilchez Carmen Lucila Infante Saavedra. "La Web del Futuro : Web Semántica". En: (2011), págs. 2-5.

[5] Juan Pablo Tendazo Rodriguez Edison Leonardo Naranjo Diaz. "Desarrollo de un Agente Buscador Inteligente de Metadatos Geográficos para la Unisig". En: (2013), págs. 3-6.

[6] José Justo Martín Romero. *http://bibing.us.es/proyectos/abreproy/4519/fichero/01-PFC+en+PDF*. 2008.

[7] Karin Becker1 - Shiva Jahangiri2 - Craig A. Knoblock. "A Quantitative Survey on the Use of the Cube Vocabulary in the Linked Open Data Cloud". En: (2015), págs. 1-11.

[8] "La Web inteligente. II Jornada de en.red.ando." En: (2001).

[9] Jirí Helmich12 - Jakub Klímek32 - Martin Necaský. "Visualizing RDF Data Cubes using the Linked Data Visualization Model". En: (2014), págs. 1-5.

[10] OPenCube Toolkit. *http://opencube-toolkit.eu/*.

[11] José Danilo Villares Pazmiño. "Las Aplicaiones OLAP y si importancia en el soporte a la toma de decisiones en los procesos de compras y ventas en la empresa DISMERO S.A, Provincia de Los Ríos." En: (2012), págs. 80, 81, 82.

[12] Kim A. Jakobsen - Alex B. Andersen - Katja Hose - Torben Bach Pedersen. “Optimizing RDF Data Cubes for Efficient Processing of Analytical Queries”. En: (2014), págs. 1-12.

[13] Xavier Sumba - Freddy Sumba - Andres Tello - Fernando Baculima - Mauricio Espinoza - Víctor Saquicela. “Detecting similar areas of knowledge using Semantic and Data Mining technologies”. En: (2016), págs. 8, 9, 10.

[14] Víctor Saquicela-Mauricio Espinoza-Fernando Baculima-José Luis Cullcay-Xavier Sumba-Freddy Sumba. “Repositorio Semántico de Investigadores del Ecuador”. En: (2015), págs. 1, 7.

[15] E. Kalampokis-A. Nikolov-P. Haase-R. Cyganiak- A. Stasiewicz- A. Karamanou-M. Zotou-D. Zeginis- E. Tambouris- K. Tarabanis. “Exploiting Linked Data Cubes with OpenCube Toolkit”. En: (2014), págs. 2, 3.

[16] Revista LOGOS CIENCIA TECNOLOGÍA. “Business intelligence y la toma de decisiones financieras”. En: (2013), pág. 127.

[17] Ora Lassila Tim Berners-Lee James Hendler. “The Semantic Web”. En: (2001), pág. 1.

[18] Konrad Hoffner - Jens Lehmann - Ricardo Usbeck. “CubeQA—Question Answering on RDF Data Cubes”. En: (2015), págs. 1-16.

[19] V. Botti V. Julián. “Agentes Inteligentes: el siguiente paso en la Inteligencia Artificial”. En: (2015), págs. 1, 3.

[20] V. Botti V. Julián. “Red de Inteligencia Compartida Organizacional como soporte a la toma de decisiones”. En: (2013), págs. 110, 113.

[21] Sucerquia Osorio Andrés Vásquez Castrillón John Bayron. “La Inteligencia de Negocios: Etapas del proceso.” En: (2011), págs. 1, 2.

[22] Yimi Alberto Acevedo Villada. “La Web Semántica y el papel del Profesional de la Inforamción frente a este proyecto”. En: (2010), pág. 2.

[23] VV. AA. *http://www.pliegosdeopinion.net/pdo1/Dossier/marco1.htm*. 2012.

[24] VV. AA. *http://www.w3.org/2000/Talks/1206-xml2k-tbl*. 2000.

[25] W3C. *http://skos.um.es/TR/rdf-sparql-query/*. 2008.

[26] W3C. *https://www.w3.org/TR/2012/WD-vocab-data-cube-20120405*.

[27] W3C. *https://www.w3.org/TR/2012/WD-vocab-data-cube-20120405/*. 2012.

[28] W3C. *https://www.w3.org/TR/ldp/*. 2015.

[29] W3C. *http://www.sidar.org/recur/desdi/traduc/es/rdf/rdfesp.htm*. 2011.

[30] W3C. *http://www.w3c.es/Divulgacion/GuiasBreves/LinkedData*.

[31] W3C. *http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica*.

[32] W3Schoo. *http://www.w3schools.com/rdf/*.